World Applied Sciences Journal 22 (1): 41-45, 2013
ISSN 1818-4952
© IDOSI Publications, 2013
DOI: 10.5829/idosi.wasj.2013.22.01.2964

# Calculating Information Entropy of Language Texts

*Bikesh Revovna Ospanova*

Karaganda State Technical University, Karaganda, Kazakhstan

**Abstract:** In the work there are shown the results of calculating information entropy of language texts. There is theoretically substantiated the use of information Shannon's formula for estimating the text perfection. The information-entropic analysis is based on the comparing of various genres and styles of the Kazakh and Russian languages. There is suggested a theoretical model of the text taking into account its hierarchic structure based on different characteristics of Russian and Kazakh language texts with one, two, three, four, five and six letter combinations defining the language hierarchic structure. In the article there are presented some aspects of the experimental approach to calculating the text entropy in Kazakh and Russian, are given experimental data demonstrating the results. Using the synergetic theory of information there was carried out the structural analysis of arbitrary texts from the side of their randomness and order by the number and frequency of individual letters.

**Key words:** Information entropy · Hierarchy · Linguistic analysis · Study · Order

## INTRODUCTION

In the modern world information presents one of the most important resources, one of the motive force of the human society development. Information processes taking place in the material world, living nature and human society are studied by almost all scientific disciplines.

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

Information and entropy characterize a complicated system from the point of view of order and chaos, at that if information is a measure of order, than entropy is a measure of chaos. This measure stretches from the maximum entropy, i.e. chaos, complete uncertainty, to the highest level of order. So, the level of order is defined by the level of information at which the system is.

In the present work we base on the terminological apparatus of information theory, in accordance with which we will understand (within the frames of the tradition laid by C. Shannon):

- "Entropy" as a measure of uncertainty (unpredictability) of the system characterized by the possibility of selecting as a following element some series of any variant of the finer number of those;
- "Information" as eliminating the system uncertainty by means of the realized selection of a variant, at that unpredictable in relation to the previous states of the selection system.

Information entropy is a measure of information randomness, uncertainty of appearing the primary alphabet symbol. If there are no information losses, it is numerically equal to the amount of information for a symbol of the transmitted message [1].

**Methodology of Studying:** In the course of studying on the complex comparative base there were used the following methods: information analysis, comparative method of revealing information-entropic characteristics of the texts, method of quantitative analysis. There were analyzed the texts of various genera-and-style nature within the frames of the text general structure aspect.

**The Main Part:** The studies devoted to methodology of entropy and information are based on the works of such scientists, as: S. Angrist and L. Hepler [2], C. Shannon [3], R. Arnheim [4], L. Whyte [5], R. Narashimha [6], R. Carnap [7] and many others.

We can cite a lot of works devoted to this subject. The concept of entropy was introduced by Clausius in the XIX century as a characteristic of the chaos extent. Using entropy it became possible to estimate such

**Corresponding Author:** Dr. Ospanova, Ul. Tereshkovoi 38-17, 100012 Karaganda, Kazakhstan.

important concepts, as order and disorder. For example, *S. Angrist* and *L. Hepler* give the following definition to entropy: "…entropy is defined as a quantitative measure of disorder in a system" [2]. In doctor of philology M.Yu. Oleshkov's opinion, "…under entropy we understand a measure of uncertainty (unpredictability) of the text in the discursive process that is characterized by the possibility of selecting as a following stage from a number of variants. The indicator of entropy characterized quantitatively the level of information order of the text as a system: the more it is, the less is the system (=text) ordered, the more its deviation from the "ideal" development. Thus, entropy is a function of state; to any state of the system can be given a certain value of entropy" [8].

There exist three variants of entropy in present day science. Let's try to define the meaning of the concept of "entropy':

- In thermodynamic (according to Clauisius) it is a function of state: entropy is proportional to the amount of associated energy in the system that cannot be converted into work;
- A measure of disorder, randomness, uniformity of molecular systems (according to Boltzmann);
- In theory of information (according to Shannon) it is a measure of credibility transmitted by the information communication channel: entropy characterizes numerically the transmitted signal credibility and is used for calculating the information amount.

Well-known physicist-mathematician S.M. Korotayev said: "It is difficult to find more general concepts for all sciences (not only natural) and sometimes more enigmatic than entropy and information". Calling the concepts universal, natural-scientific, fruitfully used in a lot of fields, the scientist expressed hope that readers will not only see the possibilities of using the entropic approach in their field but will further develop it [9].

Meanwhile, within the last decades a lot of scientists-linguists are greatly interested in the language that demonstrates the signs of a self-organizing system and is characterized as an open nonlinear system subjected to internal and external fluctuations, suggesting both instability and chaos in structures and their order. The language development is performed through its permanent instability, due to its states changing. Something new in the language appears as something sudden and unpredicted, at that emphasizing the dynamism of a natural language and objecting possible amplitudes of developing various relations of the language units. Language phenomena are considered from the point of view of complexity, nonlinearity, chaos, attractor, synergy. The language acts as a more complicated hierarchic level of self-organization and presents a macroscopic system of characters consisting of a lot of structural elements: phonemes, words, word combinations, sentences. Phonetic, grammar, syntax rules reflect the appearance of order from chaos, e. self-organization [10].

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

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 sequential realization of a certain system. The amount of information expressed by the indicated letter is absolutely equal to the entropy (uncertainty) that characterized the system of possible selections and that was eliminated as a result of selecting a certain letter.

It is known that to calculate entropy it is necessary to have a complete distribution of probabilities of possible combinations. To calculate entropy of this or that letter it is necessary to to know probabilities of occurring each possible letter. That's why entropy in linguistics is one of the most universal characteristics of the text, indicator of its complexity in theoretical-information sense.

The subject of our study is a language text. At that we emphasize our attention at a new linguistic-mathematical model for analyzing the text structure. It is built based on the fundamental law of preserving the sum of information and entropy using Shannon's formula. When characterizing in general the entropic-information (entropy is a measure of disorder, information is a measure of order) analysis of the texts we used a statistical Shannon's formula for defining the text perfection, harmony:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{11}$$

where $p_i$ is the probability of detecting any system unit in their multitude $\sum_{i=1}^{N} p_i = 1, p_i \geq 0, i = 1,2,...,N \cdot$

For linguistic a very important measure is the language entropy. It is a general measure of probabilistic-linguistic relations in the given language text. In this connection we compare the data characterizing the quantitative estimate of these measures in Kazakh and Russian.

We carried out a linguistic analysis of the texts containing 500 characters of scientific, business-official, journalistic, everyday-informal and belles-lettres styles of speech in Kazakh and Russian.

To calculate the texts information there were calculated the probabilities of occurring one letter, two-, three-, four-, five- and sox-letter combinations. In calculations there were accounted 31 letters of the Russian alphabet (letters **е** and **е**, ъ and **ь** are taken as one letter) and a blank and 43 letters (42 letters and a blank) of the Kazakh alphabet; all the rest characters (brackets, quotes, commas, etc.) were not considered. The numerical data contained in the text were presented in writing.

The calculation of probability (p) of occurring different letters in the text is achieved by the calculating a relative frequency of individual letters. To define the probability of appearing a letter in the Kazakh and Russian texts we used a classical formula of probability determination:

$$P(one\_let.) = \frac{m}{n},$$

where

$P$ is a relative frequency;

$m$ is the number of one letter occurring in the text;

$n$ is the number of occurring all the letters in the text.

As a result for Kazakh there were obtained the following values (in bits), for which it should be noted that the Kazakh alphabet contains 43 letters (42 letters, a blank), then according to Shannon's formula:

Table 1: Entropy distribution in a Kazakh text

| Entropy (E) | Scientific style of speech (SS) | Journalistic style of speech (JS) | Business-official style of speech (BOS) | Everyday informal style of speech (IS) | Belles-lettres style of speech (BS) |
|---|---|---|---|---|---|
| $H_1$ | 4,3598 | 4,4253 | 4,3443 | 4,3873 | 4,3438 |
| $H_2$ | 2,3444 | 2,7267 | 2,6006 | 2,7843 | 2,7468 |
| $H_3$ | 0,852 | 1,0687 | 1,0225 | 1,0557 | 1,2596 |
| $H_4$ | 0,2813 | 0,3301 | 0,2665 | 0,3187 | 0,414 |
| $H_5$ | 0,1882 | 0,1198 | 0,2012 | 0,1265 | 0,1091 |
| $H_6$ | 0,1657 | 0,0657 | 0,095 | 0,056 | 0,0414 |

Table 2: Entropy distribution in a Russian text

| Entropy (E) | Scientific style of speech (SS) | Journalistic style of speech (JS) | Business-official style of speech (BOS) | Everyday informal style of speech (IS) | Belles-lettres style of speech (BS) |
|---|---|---|---|---|---|
| $H_1$ | 4,364 | 4,3742 | 4,2746 | 4,3833 | 4,2758 |
| $H_2$ | 2,9766 | 3,0423 | 2,6721 | 2,8383 | 2,9952 |
| $H_3$ | 0,78243 | 0,7895 | 0,9196 | 1,0685 | 1,2651 |
| $H_4$ | 0,3426 | 0,5605 | 0,3290 | 0,3683 | 0,2986 |
| $H_5$ | 0,0615 | 0,0451 | 0,1517 | 0,0831 | 0,0749 |
| $H_6$ | 0,0537 | 0,0108 | 0,1046 | 0,0630 | 0,0252 |

Table 3: Entropy dynamics in Kazakh and Russian

|  | Kazakh |  |  |  |  | Russian |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| E | SS | JS | BOS | IS | BS | SS | JS | BOS | IS | BS |
| $H_1$ | = | > 0,05 | > 0,07 | = | > 0,06 | = | < 0,05 | < 0,07 | = | < 0,06 |
| $H_2$ | < 0,6 | < 0,3 | < 0,07 | < 0,05 | < 0,2 | > 0,6 | > 0,3 | > 0,07 | > 0,05 | > 0,2 |
| $H_3$ | > 0,6 | > 0,2 | > 0,1 | < 0,01 | = | < 0,6 | < 0,2 | < 0,1 | > 0,01 | = |
| $H_4$ | < 0,06 | < 0,2 | < 0,06 | < 0,04 | > 0,1 | > 0,06 | > 0,2 | > 0,06 | > 0,04 | < 0,1 |
| $H_5$ | > 0,12 | > 0,07 | > 0,04 | > 0,04 | > 0,03 | < 0,12 | < 0,07 | < 0,04 | < 0,04 | < 0,03 |
| $H_6$ | > 0,11 | > 0,05 | = | = | > 0,01 | < 0,11 | < 0,05 | = | = | < 0,01 |

$H_0 = \log 43 = 5{,}4$ bit.

where $H_0$ is the maximum value of the text entropy consisting in acceptance of one letter of the Kazakh text (information contained in one letter), under the condition that all the letters are considered equally probable; bit is a unit of measuring information.

Now let's consider the analysis of Russian texts in entropy distribution. As the Russian alphabet contains 32 letters (31 letters, one blank), according to this result

$H_0 = \log 32 = 5$ 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.

The calculation show that $H_{\max}$ value in the Russian language does not practically differ from $H_{\max}$ in the Kazakh language

In Russian $H_0 = \log 32 = 5$ bit
In Kazakh $H_0 = \log 43 = 5, 4$ bit

Thus, the studies showed that the dynamics of entropy $H_1$ of the texts is practically the same in scientific and informal styles of Kazakh and Russian and the text entropy, accounting $H_2$, $H_4$, for all styles in Russian is larger than in Kazakh. The text entropy of SS, JS and BOS, accounting $H_3$, $H_5$, $H_6$, in Kazakh is larger than in Russian. Alongside with this it becomes clear that in JS and BS styles of Kazakh the text entropy is larger, accounting $H_1$, $H_3$, $H_5$, $H_6$; the text entropy for all styles in Russian accounting $H_2$, $H_4$ is larger than in Kazakh within the fluctuation limits from 0.007 to 0,1. The level entropy $H_6$ is approximately equal in BOS and IS in Russian and Kazakh. These results have been obtained on the basis of the two languages analysis. There were analyzed the texts of various genres and styles. The presented indicators were analyzed using the method of calculating on, two, three, four, five and six letter combinations both in Kazakh and Russian based on the fundamental law of preserving the sum of information and entropy using Shannon's formula.

Making a conclusion we will note that on the basis of the Kazakh and Russian texts there have been obtained information characteristics of the letters that are in different positions; obtained in-letters distribution of the text entropy and given the possibility to estimate quantitatively the information ratio in the text. All this permits to come to a conclusion that information entropy can be used in any language for revealing the information distribution in the text.

From here we can conclude that the dynamics of the text information entropy decreases in the transition to a higher level, at this there increases the text information capacity which proves the language development according to the law of preserving the sum of information and entropy.

## CONCLUSIONS

Thus, the complete analysis of the study shows that the plan of building a complicated information system can be formed only at the upper hierarchic levels and to descend from there to the lower levels causing this or that order of the lements alternation.

Used by information theory statistical method of accounting inter-letter correlations in literature texts of both languages depends on the semantic context and one, two, three or more letters can be in some cases an independent word and in other cases – to enter the structure of other words.

It is obvious that the considered letter combinations relate to different hierarchic levels of the text, however, such differentiation of the levels can be performed only by the meaning that is in the analyzed text.

The reasons of occurring the studied order are always beyond the limits of competences of statistical methods. Being at the lower levels of a certain hierarchic structure, science armed with statistical methods, studies not the action of the reasons generating the studied order, but only its result. In the article the probabilistic function of entropy is used by us for the strict definition of the information and entropy amount in the texts at the word level, as using words we can make practically unlimited number of texts.

## REFERENCES

1. http://ru.wikipedia.org/wiki/Information_entropy.
2. Angrist, S.W. and L.G. Hepler 1967. Order and Chaos: Laws of Energy and Entropy. N.Y.: Basic Books, pp: 146
3. Shannon, C.E., 1948. Mathematical Theory of Communication // Bell System Technical Journal, 27: 379-423, pp: 623-656.

4. Arnheim, R., 1971. Entropy and Art. An Essay on Disorder and Order. Berkley and Los Angeles: University of California Press, pp: 61.
5. Whyte, L., 1965. Law Atomism, Structure and Form // Structure in Art and Science / Kepes Gyorgy (ed.). New York: Braziller, pp: 20-28.
6. Narashimha, R., 1994. Linguistic Entropy in Othello of Shakespeare. New Delhi: MD Publications Ltd, pp: 95.
7. Carnap, R., 1977. Two Essays on Entropy. Berkley and Los Angeles: University of California Press, pp: 115.
8. Oleshkov, M.Yu., 2006. Bases of functional linguistics: discursive aspect: Tutorial for students of Russ. Lang. and lit. Nizhni Tagil, pp: 146.
9. Korotayev, S.M., 0000. Entropy and information: universal and natural-scientific concepts. htth:// www.chronos.msu.ru/PREPORT/korotaev entropia/korotaev entropia.htm.
10. Zhanabayev, Z.Z.H., 1998. Knowledge synergy: scientific bases of estimating academic activity. Almaty: Kazakh University, pp: 37.
11. Shannon, C.E., 1963. Mathematical Theory of Communication // Works on Theory of Information and cybernetics. M.: FL, pp: 243-332.