SOME ASPECTS OF LANGUAGE MODEL IN INFORMATION THEORY

[1]Ospanova B.R., [2]Kazhikenova S.S.

[1]Karaganda State Technical University, Karaganda, e-mail: o.b.r@mail.ru;
[2]Karaganda State University n.a. E.A. Buketov, Karaganda, e-mail: sauleshka555@mail.ru

In the article there are presented some aspects of theoretic-and-experimental approach to evaluation of Russian and Kazakh texts entropy. The methodology suggested is based on the system, multilevel approach to building a complicated hierarchic system of a language.

Keywords: linguistics, entropy, information theory, ordering, scientific text, self-organization

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

When determining the quantity of information there is considered a language text which consists of letters, words, word combinations, sentences, etc. Each letter occurrence is described as a sequential realization of a certain system. The quantity of information represented by the letter indicated is equal in its absolute value to the entropy (uncertainty) which characterized the system of possible choices and which was eliminated as a result of a certain letter selection.

It's known that in order to evaluate entropy it is necessary to have a complete distribution of possible combinations probabilities. Therefore, in order to evaluate entropy of a letter it is necessary to know probabilities of every possible letter occurrence.

Research objective. Our research is conditioned by the necessity to study the text material of various genres with the aim of its improving. Any text is to be formed correctly in style, grammar, syntax, without linguistic mistakes. By means of using mathematical calculations we obtained the values of a letter entropy taking into account one, two, three, four, five or six letters of a text in the Russian and Kazakh languages.

We suggest an ideal 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.

In the general characteristic of the text entropic-information (entropy is a measure of disorder, information is a measure of eliminating disorder) analysis we used Shannon's statistical formula to determine the text perfection, harmony:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i , \qquad (1)$$

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

Before publishing Shannon's theory, Hartley suggested to determine the maximum entropy quantity by the formula:

$$H_{\max} = \log_2 N . \qquad (2)$$

Studies in the field of information theory are of a great interest. For linguistics an important measure is the language entropy. It is a general measure of probabilistic-linguistic ties in the given language text. In this connection we carry out a comparison of the data characterizing a numerical evaluation of these measures in the Kazakh and Russian languages.

As the Russian alphabet contains 32 letters (31 letters, one blank), according to this result

$$H_0 = \log 32 = 5 \text{ bits}.$$

H_0 is the maximum value of the text entropy contained in receiving 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 measuring information.

The Kazakh alphabet contains 43 letters (42 letters, one blank), so according to this result,

$$M\frac{\log 43}{\log m} = M\frac{H_0}{\log m}.$$

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

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

Here we are to note that the present day Kazakh Cyrillic alphabet is used in Kazakhstan and Mongolia. In adopted in 1940 alphabet developed by S.A. Amanzholov, there are 42 letters; 33 of them are from the Russian alphabet and 9 are specific letters of the Kazakh language: ***Ә, Ғ, Қ, Ң, Ө, Ұ, Ү, Һ, І.*** Initially the Kazakh letters were placed after the letters of the Russian alphabet, then each of them was placed after the Russian letters similar in pronunciation. The following letters: ***в, ё*** (since 1957), ***ф, х, һ, ц, ч, щ, ъ, ь, э*** are not used in purely Kazakh words. The letters ***ё, ц, ч, щ,***

ъ, ь, э are used only in the words which are borrowed from Russian or through Russian and are written in accordance with the rules of Russian orthography. The letter ***х*** in spoken speech is pronounced as ***қ***. The letter ***һ*** is used only in Arabian-Persian borrowings and is often pronounced as a dull sound ***х***. The letter ***е*** in the word absolute beginning can be pronounced as the diphthongoid [^{j}е]. The letter ***э*** is always pronounced as ***е***. The letter ***о*** in the word absolute beginning can be pronounced as the diphthongoid [^{w}о]. The letters ***і*** and ***ы*** denote sounds similar to Old Slavonic (before the reduced fall) ь и ъ. The letter ***и*** denotes pseudo-diphthongs ый, ій. The letter ***у*** denotes a non-syllabic sound similar to the Belorussian ў and pseudo-diphthongs ұу, үу, ыу, іу.

The following letters (called respectively «soft» or «narrow» and «hard» or «wide») denote the pairs of front and back vowels: ***е – а, ө – о, ү – ұ, і – ы.*** In the Arabic-Persian borrowings there is also a contraposition ***ә – а***. As the emphasis is always on the last syllable, it is not displayed in written form.

As an example there was considered a Kazakh text from scientific style of speech. The material for the experiment served an extract from the manual on music. The text contains 500 symbols with blanks and 431 symbols without blanks [3].

To calculate relative frequencies we used the formula of probability classical determination:

$$P=\frac{m}{n},$$

where n is the number of all the letters; m is the number of the letter considered.

The approximate values of individual letters frequencies in Kazakh are presented in Tables 1 and 2 (the dash denotes a blank between the words). In Table 1 the letters are placed in the alphabetic order, in Table 2 – as far as relative frequencies decrease.

Table 1

Distribution of relative frequency in the alphabetic order

Number	Letter	Relative frequency	Number	Letter	Relative frequency
1.	*blank*	0,138	*23*	*п*	0,008
2.	*а*	0,112	*24*	*р*	0,052
3.	*ә*	0,01	*25*	*с*	0,026
4.	*б*	0,018	*26*	*т*	0,042
5.	*в*	0	*27*	*у*	0,022
6.	*г*	0,004	*28*	*ұ*	0,002
7.	*ғ*	0,008	*29*	*ү*	0,008
8.	*д*	0,034	*30*	*ф*	0
9.	*е*	0,042	*31*	*х*	0,01
10	*ё*	0	*32*	*һ*	0
11	*ж*	0,014	*33*	*ц*	0
12	*з*	0,028	*34*	*ч*	0
13	*и*	0,004	*35*	*ш*	0,006
14	*й*	0,018	*36*	*щ*	0
15	*к*	0,036	*37*	*ъ*	0
16	*қ*	0,018	*38*	*ы*	0,124
17	*л*	0,036	*39*	*і*	0,032
18	*м*	0,05	*40*	*ь*	0
19	*н*	0,044	*41*	*э*	0
20	*ң*	0,026	*42*	*ю*	0
21	*о*	0,014	*43*	*я*	0,004
22	*ө*	0,01			

By equalizing these frequencies to the probabilities of corresponding letters occurrence, we'll obtain, based on Shannon's information entropy, a formula for calculating the maximum value of the text entropy accounting one letter of the Kazakh text:

$$H_1=H(\alpha_1)=b\cdot\log_a b=b\cdot\left(\frac{\ln b}{\ln a}\right);$$

$$H_1=H(\alpha_1)=-0{,}138\cdot\log_2(0{,}138)-0{,}124\log_2(0{,}124)-...-0{,}002\cdot\log_2(0{,}002)\approx 4{,}3598.$$

Table 2

Distribution of relative frequency as far as it decreases

letter rel.frequency	— 0,138	ы 0,124	а 0,112	р 0,052	м 0,05	н 0,044	е 0,042	т 0,042
letter rel.frequency	к 0,036	л 0,036	д 0,034	і 0,032	з 0,028	ң 0,026	с 0,026	у 0,022
letter rel.frequency	б 0,018	й 0,018	қ 0,018	ж 0,014	о 0,014	ә 0,01	ө 0,01	х 0,01
letter rel.frequency	ғ 0,008	п 0,008	ү 0,008	ш 0,006	г 0,004	и 0,004	я 0,004	ұ 0,002

The approximate values of frequencies of two-letter combinations in Kazakh are presented in Table 3 (the dash denotes a blank between the words). In Table 3 the letters are placed as far as relative frequencies decrease.

Table 3

Distribution of two-letter combinations relative frequencies

combination rel.frequency	ы- 0,032	-м 0,022	ры 0,022	ың 0,020	ң - 0,020	му 0,020	уз 0,020	зы 0,020
combination rel.frequency	ык 0,020	ка 0,020	ты 0,018	- т 0,018	та 0,018	н - 0,018	і - 0,016	а - 0,016
combination rel.frequency	ыр 0,016	лы 0,016	-б 0,014	ар 0,014	-ж 0,014	мы 0,014	ал 0,012	ық 0,012
combination rel.frequency	ас 0,012	сы 0,012	ба 0,012	- к 0,012	ам 0,012	ен 0,012	ер 0,012	- х 0,001
combination rel.frequency	ха 0,01	да 0,01	рі 0,01	- о 0,01	ын 0,01	нд 0,01	ан 0,01	де 0,001
combination rel.frequency	р - 0,008	кт 0,008	- ә 0,008	ән 0,008	ді 0,008	- д 0,008	п - 0,008	ай 0,008
combination rel.frequency	ны 0,008	ла 0,008	ме 0,008	жы 0,008	ні 0,006	із 0,006	жа 0,006	кө 0,006
combination rel.frequency	- а 0,006	ды 0,006	кү 0,006	үй 0,006	йл 0,006	ле 0,006	ол 0,006	ыл 0,006
combination rel.frequency	- с 0,006	рм 0,006	қ - 0,006	ор 0,004	йт 0,004	ег 0,004	ге 0,004	ім 0,004
combination rel.frequency	мі 0,004	ат 0,004	з - 0,004	зд 0,004	ағ 0,004	ға 0,004	л - 0,004	- ө 0,004
combination rel.frequency	се 0,004	ед 0,004	аң 0,004	ңа 0,004	ып 0,004	ей 0,004	рл 0,004	аш 0,004
combination rel.frequency	- е 0,004	йд 0,004	лм 0,004	ма 0,004	әр 0,002	бі 0,002	ің 0,002	ақ 0,002
combination rel.frequency	қс 0,002	өр 0,002	іп 0,002	нд 0,002	өп 0,002	ым 0,002	ыз 0,002	өт 0,002
combination rel.frequency	тк 0,002	ке 0,002	са 0,002	йы 0,002	өс 0,002	е- 0,002	тү 0,002	аб 0,002
combination rel.frequency	үс 0,002	өб 0,002	бе 0,002	йе 0,002	шт 0,002	си 0,002	ия 0,002	яқ 0,002
combination rel.frequency	еш 0,002	шқ 0,002	қа 0,002	ша 0,002	ес 0,002	ск 0,002	кі 0,002	ір 0,002
combination rel.frequency	со 0,002	то 0,002	ығ 0,002	ғы 0,002	от 0,002	ра 0,002	ад 0,002	- я 0,002
combination rel.frequency	яғ 0,002	ғн 0,002	ни 0,002	и - 0,002	он 0,002	ст 0,002	ау 0,002	у - 0,002
combination rel.frequency	бұ 0,002	ұл 0,002						

Then we'll calculate the conditional entropy $H_2 = H\alpha_1(\alpha_2)$ of the experiment α_2, consisting in determining one letter of the Kazakh text under the condition that we know the output of the experiment α_1, consisting in determining the previous letter of the same text. In accordance with the abovementioned, H_2 is determined by the following formula:

$$H_2 = H\alpha_1(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) = -0{,}032\cdot\log_2(0{,}032) - 0{,}022\cdot\log_2(0{,}022) - \ldots$$
$$-(0{,}002)\cdot\log_2(0{,}002) + 0{,}138\cdot\log_2(0{,}138) + 0{,}124\cdot\log_2(0{,}124) + \ldots + 0{,}002\cdot$$
$$\log_2(0{,}002) \approx 2{,}3444.$$

Similarly we can determine entropy H_3.

By equalizing these frequencies to the probabilities of corresponding three-letter combinations occurrence which is expressed by the difference $H_2 - H_3$, we'll obtain for three-letter entropy in Kazakh the approximate value:

$$H_3 = H\alpha_1\alpha_2(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2) = -0{,}020\cdot\log_2(0{,}020) - 0{,}020\cdot\log_2(0{,}020) - \ldots$$
$$\ldots - 0{,}002\cdot\log_2(0{,}002) + 0{,}032\cdot\log_2(0{,}032) + 0{,}022\cdot\log_2(0{,}022) + \ldots + 0{,}002\cdot\log_2(0{,}002) \approx$$
$$\approx 0{,}852.$$

The approximate values of four-letter combinations in Kazakh. By equalizing these frequencies to the probability of corresponding letters occurrence we'll obtain, based on Shannon's information entropy a formula for calculating the maximum value of the text entropy accounting four letters of the Kazakh text:

$$H_4 = H\alpha_1\alpha_2\alpha_3(\alpha_4) = H(\alpha_1\alpha_2\alpha_3\alpha_4) - H(\alpha_1\alpha_2\alpha_3) =$$
$$= -0{,}020\cdot\log_2(0{,}020) - 0{,}020\cdot\log_2(0{,}020) - \ldots$$
$$\ldots - 0{,}002\cdot\log_2(0{,}002) + 0{,}020\cdot\log_2(0{,}020) + 0{,}020\cdot\log_2(0{,}020) +$$
$$+ \ldots + 0{,}002\cdot\log_2(0{,}002) \approx 0{,}2813.$$

As a result of using the formula we'll determine entropy H_5.

Using the classical formula of determining a probability, the calculation of the entropy maximum value accounting five letters of the Kazakh text will make an approximate value:

$$H_5 = H\alpha_1\alpha_2\alpha_3\alpha_4(\alpha_5) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) - H(\alpha_1\alpha_2\alpha_3\alpha_4) =$$
$$= -0{,}020\cdot\log_2(0{,}020) - 0{,}020\cdot\log_2(0{,}020)\ldots - \ldots 0{,}002\cdot\log_2(0{,}002) + 0{,}020\cdot\log_2(0{,}0020) +$$
$$+ \ldots + 0{,}002\cdot\log_2(0{,}002) \approx 0{,}1832.$$

In accordance with the said, to determine conditional entropy H_6 there was evaluated the number of all the six-letter combinations in the text and used the formula of the classical determining of probability:

$$P = \frac{m}{n},$$

where n is the number of all six-letter combinations, m is the number of combinations, for example *music*.

$$H_6 = H\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5(\alpha_6) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6) - H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) =$$
$$= -0{,}020\cdot\log_2(0{,}020) - 0{,}020\cdot\log_2(0{,}020) - 0{,}012\cdot\log_2(0{,}012)\ldots - \ldots 0{,}002\cdot\log_2(0{,}002) +$$
$$+ 0{,}020\cdot\log_2(0{,}0020) + \ldots + 0{,}002\cdot\log_2(0{,}002) \approx 0{,}1657.$$

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

H_1	H_2	H_3	H_4	H_5	H_6
4,3598	**2,3444**	**0,852**	**0,2813**	**0,1882**	**0,1657.**

From here we can conclude that for the Kazakh language the language entropy decreases with the transition to a higher level of organization, besides, the text information capacity increases, which proves the language developing in accordance with the law of preserving the sum of information and entropy.

The calculations show that value H_{max} in Russian (the alphabet contains 32 letters (the letters е и ё, ь и ъ are expressed by the same combination and a blank (–) between the words)) does not practically differ from H_{max}

content of the Kazakh alphabet (42 letters and a blank):

$$H_0 = \log 32 = 5 \text{ bits;}$$

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

Now let’s see the analysis of the Russian text. We carried out an information-entropy analysis of an extract from the course of lectures on economic theory [4]. The extract presents a text of scientific style in which there are obvious the characteristics and signs of the language of science.

To calculate the scientific text information we counted the probabilities of occurrence of one letter, two-letter, three-letter, four-letter, five-letter and six-letter combinations in the text. When counting we took into consideration 31 letters of the Russian alphabet (letters е и е, ъ and ь were taken as one letter) and a blank, all the rest symbols (brackets, quotes, commas, etc.) were not considered. The calculations were carried out similar to the Kazakh text using Shannon’s information entropy for calculating the entropy maximum value in Russian. The text contains 500 symbols with blanks and 442 without blanks.

In order to calculate each letter relative frequency, it is necessary to divide each letter quantity by the general number of all symbols (500).

Table 4

Distribution of relative frequency in the alphabetic order

Letter	Number of the letter occurrence: number of all the letters	Relative frequency	Letter	Number of the letter occurrence: number of all the letters	Relative frequency
а	26:500	**0,052**	р	27:500	**0,054**
б	4:500	**0,008**	с	24:500	**0,048**
в	25:500	**0,05**	т	29:500	**0,058**
г	4:500	**0,008**	у	11:500	**0,022**
д	10:500	**0,02**	ф	3:500	**0,006**
е	30:500	**0,06**	х	2:500	**0,004**
ж	5:500	**0,01**	ц	1:500	**0,002**
з	10:500	**0,02**	ч	2:500	**0,004**
и	45:500	**0,09**	ш	3:500	**0,006**
й	6:500	**0,012**	щ	2:500	**0,004**
к	14:500	**0,028**	ы	6:500	**0,012**
л	18:500	**0,036**	ъ,ь	2:500	**0,004**
м	9:500	**0,018**	э	5:500	**0,01**
н	34:500	**0,068**	ю	3:500	**0,006**
о	55:500	**0,11**	я	13:500	**0,026**
п	14:500	**0,028**	space	58:500	**0,116**

Let’s place the symbols relative frequency sequentially, as far as it decreases:

Table 5

Distribution of relative frequency as far as it decreases

Letter frequency	Blank 0,116	О 0,11	И 0,09	Н 0,068	Е 0,06
letter frequency	Т 0,058	Р 0,054	А 0,052	В 0,05	С 0,048
letter frequency	Л 0,036	К 0,028	П 0,028	Я 0,026	У 0,022
letter frequency	Д 0,02	З 0,02	М 0,018	Й 0,012	Ы 0,012
letter frequency	Ж 0,01	Э 0,01	Г 0,008	Б 0,008	Ю 0,006
letter frequency	Ф 0,006	Ш 0,006	Ъ, Ь 0,004	Х 0,004	Ч 0,004
letter frequency	Щ 0,004	Ц 0,002			

As a result of our studies when counting the number of various letter combinations reiteration in a scientific text, we came to the following indications:

$$H_1 = 4,\ 364 \text{ bits};$$

$$H_2 = H_{\alpha 1}(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) = 7{,}3406 - 4{,}364 = 2{,}9766;$$

$$H_3 = H_{\alpha 1\alpha 2}(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\ \alpha_2) = 8{,}123 - 7{,}3406 = 0{,}7824;$$

$$H_4 = H_{\alpha 1\alpha 2\alpha 3}(\alpha_4) = H(\alpha_1\alpha_2\alpha_3\alpha_4) - H(\alpha_1\alpha_2\alpha_3) = 8{,}4656 - 8{,}123 = 0{,}3426;$$

$$H_5 = H_{\alpha 1\alpha 2\alpha 3\alpha 4}(\alpha_5) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) - H(\alpha_1\alpha_2\alpha_3\alpha_4) = 8{,}5271 - 8{,}4656 = 0{,}0615;$$

$$H_6 = H_{\alpha 1\alpha 2\alpha 3\alpha 4\alpha 5}(\alpha_6) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6) - H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) = 8{,}5808 - 8{,}5271 = 0{,}0537.$$

Thus, the further counting the texts from one to six-letter combinations is not similar for Kazakh and Russian. Based on the evaluations carried out it can be supposed that in scientific texts in the both languages there takes place a decrease of the uncertainty (entropy) degree with the information increase. Entropy in Kazakh and Russian is equal to (in bits):

In Kazakh

H_1	H_2	H_3	H_4	H_5	H_6
4,359	**2,344**	**0,852**	**0,281**	**0,188**	**0,165**

In Russian

H_1	H_2	H_3	H_4	H_5	H_6
4,364	**2,976**	**0,782**	**0,342**	**0,061**	**0,053**

Conclusion

Making a conclusion on this study, we would like to note that this fact is explained by a different number of the hierarchic system elements, different number of letters in the alphabets of the Russian and Kazakh languages. The text entropy decrease at the higher levels justifies the fact that for a multilevel hierarchic system it is very significant to describe a lower level as an interaction of interconnected subsystems, each of which possesses its information characteristics. We established that with transition to a higher level of the hierarchic system which is based on accounting the letter combinations, the information capacity of the texts increases. The approach considered, in our opinion, corresponds to the main requirements of the system entropy-information analysis as in the hierarchic system modeling it ensures its consideration integrity due to general theoretic and methodological conceptions.

References

1. Shannon C.E. Mathematical theory of communication // Proc. on information theory and cybernetics. – M.: IL, 1963. – P. 243–332.

2. Hartley R. Information transfer // Information theory and its applications. – M.: Fizmatgiz, 1959. – P. 5–35.

3. Orazaliyeva M.A. Manual on music. – Almaty: Almakytap baspasy, 2009. – 96 p.

4. Principles of economic theory: Course of lectures. – 2-d ed. / Gen.ed. by A.A. Kochetkov. – M.: Pub.-trad.corp. «Dashkov & Co», 2005. – 492 p.