Environment. Technology. Resources, Rezekne, Latvia
Proceedings of the 10th International Scientific and Practical Conference. Volume III, 11-16

Latvian language as a code in different communication channels

Linda Bajarune[1], Andris Ozols[2]
[1]Liepaja University,
[2]Riga Technical University

Abstract. **This paper is dedicated to analyze of Latvian language as a code in such literary communication channels like press, poet, prose, legal literature. Calculations for zero-order, first-order, second-order and third-order Shannon entropy have been made and also corresponding values of redundancy and compression coefficients have been determined. All the calculations are done with a self-made computer program. Different communication channels of Latvian language are compared mutually and also Latvian language is compared with English and Russian as codes.**

Keywords: **compression coefficient, Latvian language, redundancy, Shannon entropy.**

I INTRODUCTION

Nowadays as we speak we are saying that we live in the age of information. Information concept usually associated with the two objects - the source of information and the consumer's existence. The information is very difficult to define. Accurate and in all cases acceptable definition of information is not even created. Information processes and studies are very frequent occurrence. But Latvian language so far had not yet been analyzed from the point of view of communication theory.

Natural language is one of the main ways how to communicate. As the system of signs it is a tool to send and receive information. With its system of signs, symbols and rules of their combinations, connections and typesetting, language is a unique communication code and according to that other nonverbal codes are being used.

From its beginning language also has been a tool for information storage and makes the structure and navigation system for this stored information.

In space overloaded with information where modern technologies allows very tight interaction between people whose location is very far from each other information flows in such enormous speed. Approach how the language has been researched has to be changed and also rules for natural language have to be made than language could develop not like some abstract tool for abstract communication, but as rich and creative instrument that can be used to store and transmit information about this new and fast changing reality. That is why not only research about usage of language has been made but also it is important to pay attention to its statistic properties [1].

In 1948 C.Shannon in his article "A Mathematical Theory of Communication" developed information theory revealing the most important aspects of communication systems. The two main concepts of this theory are the concepts of probability and coding. In his theory C.Shannon introduces so called structural information. Shannon completely ignores whether the text is important, correct, incorrect or irrelevant. Similarly, questions about information senders and recipients are excluded. It is also irrelevant whether the text is logical and meaningful or letters are selected at random. Here appears a paradox - randomly selected letters provide the maximum information, whereas the text with a greater meaning and linguistic diversity corresponds to smaller information value [2].

From the point of view of communication theory, any language is a code. How informative is this code is characterized by its Shannon entropy. If diversity of language is greater, Shannon entropy also is higher.

A natural language can be considered as a complex system since the succession of its symbol units (letters, syllables or words) inside a text obeys some rules (grammatical or syntactical), which, however, are of probabilistic nature allowing the insertion of randomness in the text structure [12].

In this article, Latvian language has been studied as code of literary communication channels such as the press, poetry, prose and legal literature. It was a challenge to do something that has been never done before – to calculate the entropy of Latvian language, compare it with other languages and try to analyze it.

II MATERIALS AND METHODS

Entropy is a quantitative measure of uncertainty in thermodynamics and information theory. Entropy concept is used in various information optimal

ISSN 1691-5402

DOI: http://dx.doi.org/10.17770/etr2015vol3.182

encoding problem studies. The concept of structural information which is used in communication theory describes how much randomness is in a random event, how probable it is. In a communication theory this event is called a message. It is assumed that an ensemble of messages (e.g., letters, ciphers, pixels, etc.) is transmitted over the communication channel. These messages are determined by the code which is used. Each message i is characterized by its probability, p_i. The sum of all probabilities equals to one.

Shannon entropy is the average amount of information contained in a message Entropy is a quantity which depends only on the statistic nature of the information source expressed in message probabilities.

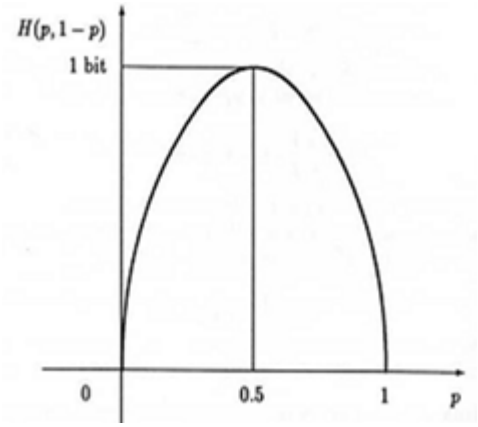

Fig.1. Entropy of two-message ensemble versus probability of one message.

In our article we consider entropy of an arbitrary message ensemble $X = (X_1, X_2, X_3, X_4....X_m.)$ which is alphabet and messages are letters.

Entropy characteristics:

1) $H(X) \geq 0$
2) Entropy is additive for two independent message ensembles X and Y
   $H(X + Y) = H(X) + H(Y)$,
3) Entropy is a limited quantity:
   $H \leq H_{max} = \log_2 m$.

We shall consider ensembles as Markov sources where the probability of each message depends on the appearance of previous messages. The appearance of a certain letter in the text generally depends on the previous letters .Thus we have to use the conditional entropies of different orders.

Zero-order entropy $H(0)$ does not take into account any interdependence of messages (they are assumed to be independent and with equal probability $1/m$) and it is calculated by the following formula

$$H(0) = \log_2 m \quad \text{bits/symbol,} \qquad (1)$$

where m is the number of letters in the alphabet and also the space between words, because entropy is being calculated to the written language. It can be shown that $H(0)=H_{max}$.

First-order entropy $H(1)$ also does not take into account the interdependence of messages, but in this case their probabilities are not equal:

$$H(1) = \sum_{i=1}^{m} p_i \log_2 p_i \quad \text{bits/symbol,} \qquad (2)$$

where p_i – probability of the message I (letter). The first-order entropy dependence on the message probability is shown in Fig.1 for the special case of the ensemble of two messages with probabilities p and $1-p$**.** It is seen that entropy maximum is achieved in the case of equal message probabilities, i.e., in this case $H(1)=H(0)=H_{max}$ as stated before.

Second-order entropy is calculated by formula

$$H(2) = \sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_{j|i} \log_2 p_{j|i} \quad \text{bits/symbol,} \qquad (3)$$

where pj | i - j conditional probability of the letter, if the former has been the letter i. To calculate the second-order entropy must take into account what the symbol stands before the symbol, or all possible combinations of the two symbols.

Third order entropy is calculated by formula

$$H(3) = -\sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_{j|i} \sum_{k=1}^{m} p_{k|j,i} \log_2 p_{k|j,i} \qquad (4)$$

It is possible to calculate and report the source of redundancy, if entropy has been calculated.

$$\rho = 1 - \frac{H(A)}{H_{\max}(A)} = 1 - \frac{H_n(A)}{H_0(A)} \qquad (5)$$

With information redundancy understands the duplicate or collateral data activation in system data blocks which the withdrawal does not detract from the adequacy of an array of real objects they describe. [5] Redundancy in information theory is the number of bits that are used to send the message minus the fair amount of information in bits. Data compression is a way to exclude unwanted redundancy, but if the message shall be carried out in a noisy channel with a limited capacity, then the redundancy is desirable.

So redundancy in our language is the words that we say, but even without these words, the information is comprehensible. About that we can make sure every time when we send text messages, trying to say as much as possible with the least possible symbols to convene only one message - words, letters are omitted or even written without spaces, but the text still is understandable.

Txt wtht vwls. TextWithoutSpacing. –if we can read and still understand this text, which is written without vowels and without spacing, that means that vowels or spaces are redundant for this message.

But redundant words or letters are because we can understand the information in a noisy place. For example, where is salt package? If in the moment when that being said drives noisy large car, word salt

can sound like some other word, than redundant word – package – will help to understand the real meaning of the sentence. This example is very simple, but in most of cases if in the sentence one word is missing, we can guess this word. If one letter is missing, clearly we can guess this letter.

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Or rather...According to a researcher (sic) at Cambridge University, it doesn't matter in what order the letters in a word are, the only important thing is that the first and last letter be at the right place. The rest can be a total mess and you can still read it without problem. This is because the human mind does not read every letter by itself but the word as a whole [6].

If 50% of language is redundant than it is possible to save 50% investment what needs to send electronic message in this language. Something similar happens when the file is compressed. If only there is noise somewhere in transmission and one of the symbols from compressed file is destroyed, then it is not possible to repair original file.

In the digital coding redundancy plays an important role in using encryption with numbers of even ones. For letter A in binary system stands 01000001. So, to transmit letter A we need 8 bits to send these 8 symbols. But if the line is with interferences and we receive combination with mistake – 010000?1, we cannot tell anymore which this letter is. It can be A if the missing symbol is 0, but it can be C if the missing symbol is 1. Of course in the normal context would not have problems to understand it, but if redundancy would be already used and file is compressed? In that case we should add parity bit. That would be another redundant bit, but with it would be possible to solve the problem. If the sum of numbers is even, than 0 is added, but if odd, than number 1 is received. So if we receive 010000?10, added 0 tells us that we should receive A (0100001), but if we receive 010000?11, then added number 1 tells us that we should receive letter C (01000011).

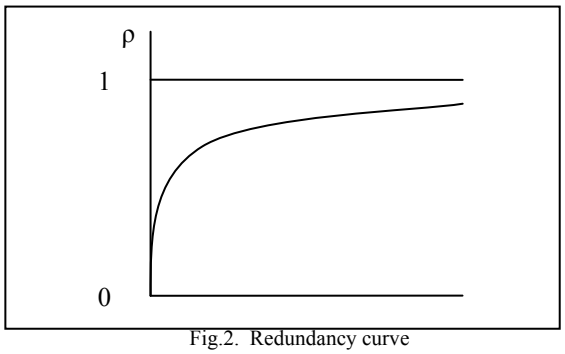

Fig.2. Redundancy curve

If really noisy channel is expected, it is possible to come to an agreement to send parity bits after every 4 bits. It could seem unnecessary to send redundant bits, which do not need. But if we are compressing text form 10 000 to 8000 symbols, excluding redundant symbols, for every transmitted sign we should add also parity bite – it would be 8000 parity bits. 8000 parity bits are equivalent 1000 symbols, which mean that it is more profitable [7].

Redundancy is closely related to compression coefficient. Figure Nr.2 shows the curve of redundancy. It calculates by formula:

$$r = \frac{H(A)}{H_{\max}(A)} = \frac{H_n(A)}{H_0(A)} = 1 - \rho \qquad (6)$$

In language as a message can consider:

1) Letter
2) Word
3) Sentence

In Latvian language sound conforms to letter, that is why letters are considered as code combination, which conforms to the message- sound.

Letters are the basic of all language, so the letters are chosen as code combination. Language is code which is used to transmit information, and letters are code combinations for information coding.

In this paper first, second and third order entropy results for three information channels - press, poet, prose, legal literature. - are given.

To make calculations computer program was made. Program was based on web page, with php coding language, and also little bit of html language. All the calculations was made with php script and MySQL data base management system [8].

The principle how program is working

1) Analyzed text is written or copied in the input field. After adding text, have to press button "Add"
2) Program replaces all capital letters with small letters, takes off all the punctuation and replaces spacing with symbol „_"
3) Text is processed forming combinations of three letters, two letters and one letter. At first in the data base program writes all the new combinations, but if the combination is found repeatedly, then the number of combination is increasing by one.
4) When analysis of the text is done, the number of combinations is placed in formulas to calculate first, second and third order of entropy. Results H(1), H(2) and H(3) are shown on the screen.

III RESULTS AND DISCUSSION

Fig.3 shows average results of entropy from all analyzed .texts.

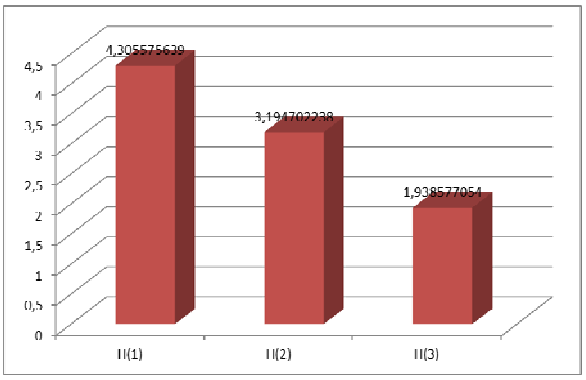

Fig.3. Average entropy in bits.

Table 1 shows entropy of Latvian language calculated by the author of this paper, entropy of English language calculated by C.E.Shannon and entropy of Russian language given by professor A.Ozols in his lectures [9]. Entropy of Latvian language is the largest, because Latvian language has more letters in the alphabet, but the value of entropy decrease faster, because combinations repeats more rarely and the compression also is the greatest.

TABLE I

COMPARISON OF ENTROPY VALUES

H(N)	Latvian	English	Russian
H(0)	5.0875	4.75	5
H(1)	4.3056	4.07	4.05
H(2)	3.1947	3.36	3.52
H(3)	1.9386	2.77	

Using values of entropy it is possible also to calculate redundancy and compression coefficient.

$$\rho(H1)=1-\frac{H(A)}{H_{\max}(A)}=1-\frac{H_n(A)}{H_0(A)}=1-\frac{4,3056}{5,0875}$$

$$=0,1537=15,4\%$$

$$\rho(H2)=1-\frac{H(A)}{H_{\max}(A)}=1-\frac{H_n(A)}{H_0(A)}=1-\frac{3,1947}{5,0875}$$

$$=0,3720=37,2\%$$

$$\rho(H3)=1-\frac{H(A)}{H_{\max}(A)}=1-\frac{H_n(A)}{H_0(A)}=1-\frac{1,9386}{5,0875}$$

$$=0,6189=61,9\%$$

Redundancies corresponding to the entropies of three orders are shown in Figure nr.4.

Comparing redundancy of Latvian language with English language and Russian language, Latvian language has the highest redundancy. It is because written Latvian language is understandable also without signs of lengthening, cedillas and without vowels. Higher entropy is possible to achieve if diversity is higher, which means combinations repeats differently, also that increases the value of entropy. If the most frequently used combination in text is dropped out, it would be possible to understand the text anyway.

TABLE II

VALUES OF REDUNDANCY

H(N)	Latvian	English	Russian
H(1)	15,4%	14,3%	19%
H(2)	37,2%	29,2%	29,6%
H(3)	61,9%	41,7%	

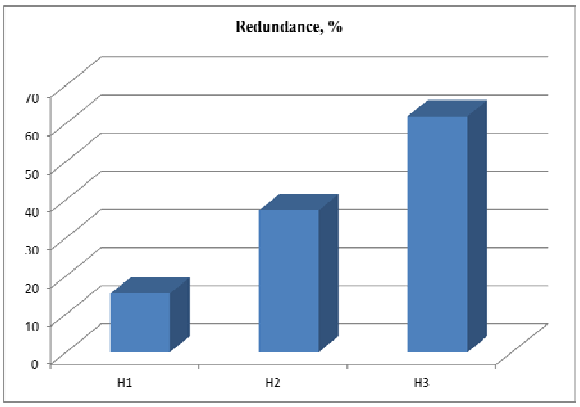

Fig.4. Redundancies corresponding to the entropies of three orders.

To define what would be the most exact redundancy, at which order of entropy it should be calculated, authors had to calculate average length of word. That is 4.52 and was calculated manually from one of the chosen texts of prose. Rounding up it shows that 5th order of entropy should be calculated. Text breakdown by word lengths, see Figure No.5.

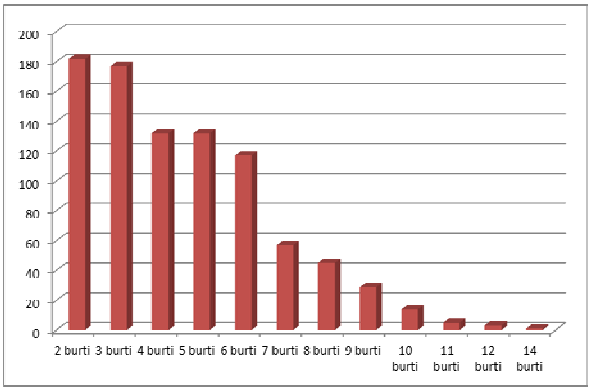

Fig.5. Word frequency versus word length.

Compression coefficient results:

At H(1) $r=\frac{H(A)}{H_{\max}(A)}=\frac{H_n(A)}{H_0(A)}=1-\rho=0,8463$

At H(2) $r=\frac{H(A)}{H_{\max}(A)}=\frac{H_n(A)}{H_0(A)}=1-\rho=0,628$

At H(3) $r=\frac{H(A)}{H_{\max}(A)}=\frac{H_n(A)}{H_0(A)}=1-\rho=0,3811$

Results also are displayed in Figure nr.6.

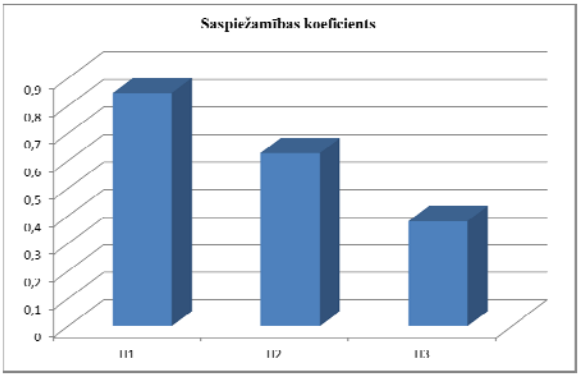

Fig.6. Compression coefficient

Average values of entropy in different information channels are summarized in Figure nr.7. First order entropy shows that the highest entropy is for press channel, followed by legal literature, then prose and the last poetry. However, the second order entropy the smallest is for legal literature but the highest for press and prose. Second-order entropy shows one letter dependence before standing letter. The higher the entropy, the greater is the appearance probability for two-letter combination. The third order entropy calculates sum of probabilities, if the appearance of one letter is dependent on before standing two letters. The smallest is the third order entropy for legal literature, followed by poetry, but the highest entropy is for prose.

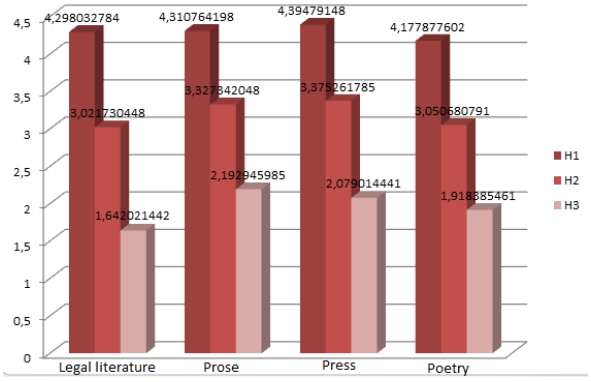

Fig.7. Average values of entropy in different information channels

IV CONCLUSIONS

To calculate the most precise results entropy of Latvian language was calculated from different information sources. All the calculations were made with web based computer program which was tested and entropy for one sentence was calculated manually.

Entropy till the third order was calculated and the results were compared for different information channels and analyzed. After these results value of average entropy was determined.

TABLE III
AVERAGE ENTROPY

H(N)	Latvian language
H(0)	5.0875
H(1)	4.3056
H(2)	3.1947
H(3)	1.9386

By defining average length of words, author made conclusion that to calculate the most precise redundancy of Latvian language, results for fifth-order entropy would need. We also can admit from the calculated average length of word that value of fifth-order entropy would be the sufficient exact to characterize Latvian language.

Comparing the entropy of the Latvian language with other languages, you can see diversity of the results.

Trend can be observed that when entropy order increases, entropy value descends slower. It can be observed due to the estimated H8 in Russian language. With the increase of entropy order ties between the letters of the word also decreases.

The probability that combinations with such a large number (more than five) will appear more than once in order not to give the entropy value equal to zero is very small. Except in cases where in the text is discussed specific topic where one word with a large number of symbols repeats (such as the recipe book of potato dishes). Calculating entropy after H5 should be interesting as a message to choose the word. But in that case definitely very long and different texts should be selected.

Entropy has an unpredictable value in its every order. However, it can be analyzed and explained. Entropy can be used as characteristics of a text both by doing calculations and making experiments. Entropy can characterize the source of information – type of a text or an author.

The larger the variety is in the text, the higher the entropy. Entropy is not a constant value. It can change by time because people use new words. Complexity and variety of the language increases consequently. According to the results, entropy changes in different sources.

It can be concluded that colloquial speech would reach the higher entropy value. The assumption is further reinforced with calculated entropy for one of the prose works. Entropy value was at its highest, because colloquial speech was used.

V REFERENCES

[1] C.E.Shannon “A Mathematical Theory of Communication”, The Bell System Technical Journal,Vol. 27, pp. 379–423, 623–656, July, October, 1948

[2] W. Gitt “Information, Science and Biology” Journal of Creation 10(2):181-187, August 1996. Available: https://answersingenesis.org/genetics/information-theory/information-science-and-biology/ [Acessed Jan 20, 2015]

[3] The Electronic Frontier Foundation “Data compression home page” [Onine]. Available: http://www.data-compression.com/index.shtml [Accessed: Jan. 20, 2015]

[4] S.Haykin “Communication Systems” 4th edition – USA: McMaster University, „John Wiley & Sons, inc.”, pp.816, 2001.

[5] Latvian Academy of Sciences “Terminology Commission data base AkadTerm” [Online] Availabe http://termini.lza.lv/term.php?term=redundance&list=&lang=LV&h=yes [Accessed: Jan. 20, 2015]

[6] Keith Rayner, Sarah J. White, Rebecca L. Johnson, and Simon P., Raeding Wrods With Jubmled Lettres There Is a Cost Liversedge, *Psychological Science*, 2003, 17(3), 192-193

[7] University of Texas at Austin “Compuer science Home page” [Onine] Available https://www.cs.utexas.edu/~eberlein/cs337/errorDetection3.pdf [Acessed Jan 20, 2015]

[8] Society Homo Culturalis “HC.lv - portal" [Online] Avaiable http://web.hc.lv/kods/php-mysql/raksti/izstrades-vides-sagatavosana-uz-windows/ [Acessed Dec.10, 2008]

[9] Ozols A., “Signal transmission theory” Lecture synopsis, Riga Technical university, Latvia, 2009.

[10] “Datu kompresija kļūst arvien aktuālāka”, *Sakaru pasaule* , Riga, 2(50),2008.

[11] .R.Ospanova.”Calculating Information Entropy of Language Texts”, World Applied Sciences Journal 22 (1): 41-45, 2013.

[12] C. Papadimitriou, K. Karamanosa, F.K. Diakonos, V. Constantoudis, H. Papageorgiou. “Entropy analysis of natural language written texts, Contents lists available at ScienceDirect Physica A.Greece, 2010 [Available: journal homepage]: www.elsevier.com/locate/physa