Bulletin KRASEC. Phys. & Math. Sci. 2016. V. 13. no. 2. pp. 62-66. ISSN 2313-0156

DOI: 10.18454/2313-0156-2016-13-2-62-66

INFORMATION AND COMPUTATION TECHNOLOGIES

MSC 54C70

CALCULATION OF KARACHAY-BALKAR TEXT ENTROPY AND PHRAS SIMULATION

M. B. Tkhamokov, A. L. Nagorov, Z. O. Beslaneev, A. Kh. Kodzokov

Kabardino-Balkarian state university of H.M. Berbekov 360004, KBR, Nalchik, Chernyshevsky str., 173

E-mail: kidmus@mail.ru

The article attempts to estimate the entropy of Karachay-Balkar printed texts. Works of famous national writers and texts of periodicals were considered as the subjects of investigation. The repetition frequency of letters, different combinations were estimated and phrases were modeled on the basis of the obtained results by a program written by the authors. Standard methods were used to calculate the characteristics. The entropy value to the twenty-fifth order and the value of the language redundancy were obtained. The results of studies of national and foreign authors in the field of entropy calculation are shown. The entropy orders of different European languages are compared.

Key words: entropy, Karachay-Balkarian alphabet, probability, redundancy

Introduction

It is known [1, p. 237] that to send a M – letter message (where is considered to be large enouth) over a communication link admitting m different elementary signals, one needs $\frac{MLogn}{Logm}$ sygnals, where n is the number of letters of an «alphabet», which are used to write a message. As long as the Karachay-Balkar «telegraph» alphabet contains 32 letters (here we do not separate the letters е and ё, ь and ъ, which are written by the same combination of elementary signals in most of telegraph codes and consider a «zero letter», the space between words, to be a letter), then according to this result, to send a M-letter message, one needs to use $\frac{MLog32}{Logm} = \frac{MH_0}{Logm}$ elementary signals. Here $H_0 = Log_2 32 = 5$ is the entropy of the experience which is the usage of

Tkhamokov Muslim Bashirovich – Senior Lecturer, Department of Computational Mathematics, Kabardino-Balkarian state university of H.M. Berbekov, Nalchik, Republic of Kabardino-Balkaria, Russia.

Nagorov Aslan L'vovich – High teacher of IMOAS chair, Kabardino-Balkarian state university of H.M. Berbekov, Nalchik, Republic of Kabardino-Balkaria, Russia.

Beslaneev Zalimbek Olegovich – High teacher of IMOAS chair, Kabardino-Balkarian state university of H.M. Berbekov, Nalchik, Republic of Kabardino-Balkaria, Russia.

Kodzokov Azamat Khasanovich – Senior Lecturer, Dep. of Mathematical Analysis and Function Theory, Kabardino-Balkarian state university of H.M. Berbekov, Nalchik, Republic of KabardinoBalkaria, Russia.

one letter of Karachay-Balkar text (information contained in one letter) under the condition that all the letters are considered to be equally probable. However, in fact, the occurrence of different letters in a message in the Karachay-Balkar language is far from equally probable. To obtain a text in which each letter contains 5 bit of information, we cannot just take a fragment from some book in Balkar language. To do that we need to write down 32 letters on separate sheets of papers, to put these sheets into a box and then to pull them out one at a time writing them down and putting the sheets back into a box and mixing them all again. When making such an experiment, we come to a «phrase» such a follows:

пспеи хмревф дквддяиъсчюцшзмфоофвэшкю

тбйэзблзиеюиемщвъзъпъбвфпючфпючфхюуаакдцвчтэфйгеждчъзшврпючржжес.

Though the text is compound of the letters of the Balkar alphabet, it has little to do with the Balkar language!

For more accurate estimation of information contained in one letter of a Balkar text, we need to know the probabilities of occurrence of different letters. These probabilities may be determined by taking a small fragment in the Balkar language and calculating relative frequencies of separate letters for it. Strictly speaking, these letters may somewhat depend on the character of a text. Thus, for a reliable determination of an «average frequency» of a letter, it is better to have several different texts borrowed from different sources.

As a text for the investigation, we used different sources: a book by Musukaeva S.A. « КЪАРАЧАЙ-МАЛКЪАР ХАЛКЪ ЖОМАКЪЛА », articles from a newspaper «Zaman» and from a journal «Mingi tau».

Method of the investigation

The investigation consists in the direct calculation of H_0 and H_1, the entropies of zero and first orders of approximation, and determination of the upper estimates H_n for the entropies of n approximation order. Graphemes of the Balkar language were decomposed. Thus, we considered that the alphabet of the text contained 32 letters (31 letters of the Russian language and a space). Thus, H_0 turned out to be equal to $\log_2 32 = 5$. H_1 was calculated in the usual way by table 1 compounded on the basis of the investigation of the text mentioned above.

Table 1

letter	-	А	Л	Н	Е,Ё	И	Ы	Р
rel. frequency	0,141	0,128	0,066	0,063	0,054	0,053	0,049	0,042
letter	У	Д	К	Т	Ъ,Ь	Г	Б	С
rel. frequency	0,040	0,037	0,035	0,033	0,031	0,030	0,027	0,022
letter	М	Й	Ю	О	З	П	Х	Ж
rel. frequency	0,021	0,016	0,016	0,015	0,014	0,015	0,012	0,012
letter	Ш	Э	Ч	Я	Ф	В	Ц	Щ
rel. frequency	0,010	0,010	0,007	0,002	0,005	0,000	0,000	0,000

Having equated these frequencies to the probabilities of occurrence of corresponding letters, we obtain an approximate value for the entropy of one letter of a Balkar text

$$H_1 = -0,141\log 0,141 - 0,128\log 0,128 - 0,066\log 0,066 - \ldots - 0,001\log 0,001 \sim 4,168024002.$$

It is clear from the comparison of this value with $H_0 = Log_2 32 = 5$ that inequality of occurrence of different latters of the alphabet results in the reduction of the information which is contained in one letter of a Balkar text by about 0,831975998 bit.

Applying this circumstance, we can decrease the number of elementary signals necessary to transmit a – letter message to the value $M\frac{H_1}{Logm}$(i.e. in the case of a binary code, to the value $H_1 M \approx 4,463793204$). The value of the average number of elementary signals equal to $\frac{H_1}{Logm}$, falling within one letter of a transmitted message, is not the best. In fact, when determining the entropy $H_1 = H(\alpha_1)$ of an experiment α_1, which was to determine one letter of a Balkar text, we considered all the letters to be independent. It means that to compose a «text» in which each letter contains $H_1 = 4,168024022$ bits of information, we should use a box with thoroughly mixed 1000 sheets, 141 of which are blank, letter А is written on 128 of them, letter Л is written on 66, . . . , at last, letter Ф is written on 1 sheet. Pulling out the sheets one at a time from such a box, we come to a «phrase» such a follows:

лр ег таатеи ыхзаалыптаалйлйурск гаеъее агс ъш заае ууаашаии ршзл алгсианм нл скбтбанеюлыкзъълха уры.

This «phrase» resemble more the meaningful Balkar speech than the previous one (comparatively probable distribution of the number of vowels and consonants and the average length of a «word» are observed). However, it is still far from a meaningful text.

The discrepancy of our phrase with a meaningful text is explained by the fact that actually the successive letters of Balkar text are not independent.

The presence of additional regularities which were not taken into account in our «phrase» results in decrease of the degree of uncertainty (entropy) of one letter from the Balkar alphabet. Thus, when sending such a text over a communication line, we may decrease the average number of elementary signals for one letter. We just need to estimate the conditional entropy $H_2 = H_{\alpha_1}(\alpha_2)$ of experiment α_2. It consists in the determination of a letter of a Balkar text under the condition that we know the result of the experiment α_1 involving the determination of a preceding letter of the same text (we should note that when receiving a consecutive letter, we always know the preceding letter). The conditional entropy H_2 is determined by the following formula:

$$H_2 = H_{\alpha_1}(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) = -p(--)\log p(--) - p(-a)\log p(-a) - \ldots$$

. . . - p(яя)log(яя)+p(-)logp(-)+. . . +p(я)log(я).

Having calculated these values by a program, we obtained the following results:

$$H_2 = H_{\alpha_1}(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) = 3,474717828,$$

$$H_3 = H_{\alpha_1\alpha_2}(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2) = 2,5452550748,$$

$$H_4 = H_{\alpha_1\alpha_2\alpha_3}(\alpha_4) = H(\alpha_1\alpha_2\alpha_3\alpha_4) - H(\alpha_1\alpha_2\alpha_3) = 1,5699993722,$$

$$H_5 = H_{\alpha_1\alpha_2\alpha_3\alpha_4}(\alpha_5) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) - H(\alpha_1\alpha_2\alpha_3\alpha_4) = 1,0138214113,$$

. . . . . ,

$$H_{25} = H_{\alpha_1\alpha_2\alpha_3\ldots\alpha_{24}}(\alpha_{25}) = H(\alpha_1\alpha_2\ldots\alpha_{25}) - H(\alpha_1\alpha_2\ldots\alpha_{24}) = 0,0019254910.$$

If we know the value H_2, we can make an experiment and obtain the following result:
лататодайы болалыдып къармни аханинген келауаннды ай шчы деры.

If we know the value H_3, we can make an experiment and obtain the following result:
дермекди мемюрегенг бол хшюн ал жону да бла салдюрлик.

The sounding of this «phrase» is closer to the Balkar language than the phrases written in the first and in the second cases.

For H_5, the modelling brought about the following result:
жетгендиле бошады да да халкъ берсенг да къоюп келген.

Discussion the investigation results

The average number of elementary signals necessary to send one letter of a text cannot be less than $\frac{H_\infty}{\log m}$. On the other hand, coding is possible when this average number is arbitrary close to the value $\frac{H_\infty}{\log m}$. The difference $1 - \frac{H_\infty}{H_0}$ showning how much the relation of «limit entropy» H_∞ to the value $H_0 = \log n$ characterizing the greatest information which may be contained in one letter of an alphabet with the given number of letters is less than a unit was called redundancy of a language by Shennon. In our case we have the following result:

$$R = 1 - \frac{H_{25}}{H_0} = 0,999614901.$$

Such a redundancy allows us to crunch a telegraph text suppressing some easily recoverable words (prepositions and conjunctions). It also allow us to recover easily the initial text even if there are many mistakes in a telegram or misprints in a book.

Redundancy R is quite an important characteristic of a language. To compare the results obtained for the Balkar language, we give the entropy values for some European languages:

Table 2

language	English	German	French	Spanish	Balkar
H_1	4,03	4,10	3,96	3,98	4,17

For the English language, Shennon obtained the following entropy values:

Таблица 3

H_0	H_1	H_2	H_3	H_5	H_8
4,76	4,03	3,32	3,10	˜2,1	˜1,9

We obtained the following results for the Balkar language:

Таблица 4

H_0	H_1	H_2	H_3	H_5	H_8
5	4,1680	3,2883	2,7266	1,6285	0,5713

Shennon's experiments [2, p. 669] showed that the value H_{100} is apparently between 0,6 and 1,3 bit. For the English language, the redundancy is about 80%. For the German language, Kupfmuller [3, pp. 265-272] obtained the value 70%. For the French language, the redundancy was calculated by N.V. Petrova [4, pp. 130-152] and it was about 71%.

References

1. Yaglom A. M., Yaglom I. M. Veroyatnost' i informatsiya [Probability and information]. Moscow. Nauka. 1973. 512 p.

2. Shennon K. Raboty po teorii informatsii i kibernetike [Works in information theory and Cybernetics]. Moscow. Publishing house of foreign literature. 1963. 830 p.

3. Kupfmuller K. Entropiya nemetskogo yazyka [Entropy of the German language]. FTZ. no. 6. 1954. pp. 265 – 272.

4. Petrova N. V. Entropiya frantsuzskogo pechatnogo teksta [The entropy of printed French]. Izvestiya Akademii nauk SSSR. Seriya literatury i yazyka. vol. 24. no. 1. 1965. pp. 63–67.

For citation: Tkhamokov M. B., Nagorov A. L., Beslaneev Z. O., Kodzokov A. Kh. Calculation of Karachay-Balkar text entropy and phras simulation. *Bulletin KRASEC. Physical and Mathematical Sciences* 2016, vol. **13**, no **2**, 62-66. DOI: 10.18454/2313-0156-2016-13-2-62-66

Original article submitted: 29.03.2016