Вестник КРАУНЦ. Физ.-мат. науки. 2016. № 2(13). С. 68-72. ISSN 2079-6641

DOI: 10.18454/2079-6641-2016-13-2-68-72

ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ

УДК 519.722

ПОДСЧЕТ ЭНТРОПИИ КАРАЧАЕВО-БАЛКАРСКИХ ТЕКСТОВ И МОДЕЛИРОВАНИЕ ФРАЗ

М. Б. Тхамоков, А. Л. Нагоров, З. О. Бесланеев, А. Х. Кодзоков

Кабардино-Балкарский государственный университет им. Х.М. Бербекова
360004, КБР, г. Нальчик, ул. Чернышевского, 173

E-mail: kidmus@mail.ru

В этой статье сделана попытка оценить энтропию карачаево-балкарских печатных текстов. В качестве исследуемого объекта были взяты произведения известных национальных писателей, а также тексты периодических изданий. С помощью программы, написанной авторами, проведены расчеты частотности букв, различных комбинаций, а также смоделированы фразы на основе полученных результатов. При подсчете характеристик использовались известные стандартные методики. Получено значение энтропии до двадцать пятого порядка и значение избыточности языка. Приведены результаты исследований отечественных и иностранных авторов в области подсчета энтропии. Проведен сравнительный анализ порядков энтропии для различных европейских языков.

Ключевые слова: энтропия, карачаево-балкарский алфавит, вероятность, избыточность

INFORMATION AND COMPUTATION TECHNOLOGIES

MSC 54C70

CALCULATION OF ENTROPY KARACHAY-BALKAR TEXTS AND SIMULATION OF PHRASES

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

This article attempts to estimate the entropy Karachay-Balkar printed texts. As of the object were taken by famous national writers, as well as the texts of periodicals. With this program, written by the authors calculated the frequency of the letters, different combinations and phrases modeled on the basis of the results obtained. When calculating performance used known standard techniques. An entropy to the twenty-fifth day of the order and the value of the redundancy of the language. The results of studies of domestic and foreign authors in the field of counting entropy. A comparative analysis of the different orders of the entropy for European languages.

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

Введение

Известно [1, с. 237], что для передачи M – буквенного сообщения (где считается достаточно большим) по линии связи, допускающей m различных элементарных сигналов, требуется затратить $\frac{MLogn}{Logm}$ сигналов, где n – число букв «алфавита», с помощью которого записано сообщение. Так как карачаево-балкарский «телеграфный» алфавит содержит 32 буквы (мы здесь не различаем буквы е и ё, ь и ъ, которые в большинстве телеграфных кодов передаются одной и той же комбинацией элементарных сигналов, но причисляем к числу букв и «нулевую букву» – пустой промежуток между словами), то согласно этому результату на передачу M – буквенного сообщения надо затратить $\frac{MLog32}{Logm} = \frac{MH_0}{Logm}$ элементарных сигналов. Здесь $H_0 = Log_2 32 = 5$ – энтропия опыта, заключающегося в приеме одной буквы карачаево-балкарского текста (информация содержащаяся в одной букве), при условии, что все буквы считаются одинаково вероятными. На самом деле, однако, появление в сообщении на карачаево-балкарском языке разных букв совсем не одинаково вероятно. Для получения текста, в котором каждая буква содержит 5 бит информации, нельзя просто взять отрывок из какой-либо книги на балкарском языке; для этого требуется выписать 32 буквы на отдельных билетиках, сложить все эти билеты в урну и затем вытаскивать их по одному, каждый раз записывая вытянутую букву, а билетик, возвращая обратно в урну, и снова перемешивая ее содержимое. Произведя такой опыт, мы придем к «фразе» вроде следующей:

пспеи хмревф дквддяиъсчюцшзмфоофвэшкю

тбйэзблзиеюиемщвъзъпъбвфпючфпючфхюуаакдцвчтэфйгеждчъзшврпючржжес.

Этот текст, хоть он и составлен из букв балкарского алфавита, имеет мало общего с балкарским языком!

Для более точного вычисления информации, содержащейся в одной букве балкарского текста, надо знать вероятности появления различных букв. Эти вероятности можно определить, взяв достаточно большой отрывок, написанный на балкарском языке, и рассчитав для него относительные частоты отдельных букв. Строго говоря, эти частоты могут несколько зависеть от характера текста; поэтому для надежного определения «средней частоты» буквы желательно иметь набор различных текстов, заимствованных из различных источников.

В качестве исследуемого текста были использованы различные источники: книга Мусукаевой С.А. « КЪАРАЧАЙ-МАЛКЪАР ХАЛКЪ ЖОМАКЪЛА », статьи газеты «Заман» и журнала «Минги тау».

Методика исследования

Исследование состояло в непосредственном подсчете H_0 и H_1 – энтропий нулевого и первого порядка приближения – и нахождения верхних оценок H_n для энтропий порядка приближения n. При этом графемы балкарского языка разлагались на составные элементы. Таким образом, считалось, что алфавит, с помощью которого составлен текст, содержит 32 буквы (31 буква русского языка, и пробел). Поэтому H_0 оказалось равным $\log_2 32 = 5$. H_1 подсчитывалась обычным образом с помощью таблицы 1, составленной на основе исследования указанного выше текста.

Таблица 1

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

Приравняв эти частоты вероятностям появления соответствующих букв, получим для энтропии одной буквы балкарского текста приближенное значение:

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

Из сравнения этого значения с величиной $H_0 = Log_2 32 = 5$ видно, что неравномерность появления различных букв алфавита приводит к уменьшению информации, содержащейся в одной букве балкарского текста, примерно на 0,831975998 бит.

Воспользовавшись этим обстоятельством, можно уменьшить число элементарных сигналов, необходимых для передачи – буквенного сообщения до значения $M\frac{H_1}{Logm}$(т.е. в случае двоичного кода – до значения $H_1 M \approx 4,463793204$). Но и равное $\frac{H_1}{Logm}$ значение среднего числа элементарных сигналов, приходящихся на одну букву передаваемого сообщения, также не является наилучшим. В самом деле, при определении энтропии $H_1 = H(\alpha_1)$ опыта α_1, состоящего в определении одной буквы балкарского текста, мы считали все буквы независимыми. Это значит, что для составления «текста», в котором каждая буква содержит $H_1 = 4,168024022$ бит информации, мы должны прибегнуть к помощи урны, в которой лежат тщательно перемешанные 1000 бумажек, на 141 которых не написано ничего, на 128 – написана буква А, на 66 – Л,. . . , наконец, на 1 бумажке – буква Ф. Извлекая из такой урны бумажки по одной, придем к «фразе» вроде следующей:

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

Эта «фраза» несколько более похожа на осмысленную балкарскую речь, чем предыдущая (здесь все же наблюдается сравнительно правдоподобное распределение числа гласных и согласных и близкая к обычной средняя длина «слова»), – но и она, разумеется, еще очень далека от разумного текста.

Несходство нашей фразы с осмысленным текстом естественно объясняется тем, что на самом деле последовательные буквы балкарского текста вовсе не независимы друг от друга.

Наличие в балкарском языке дополнительных закономерностей, не учтенных в нашей «фразе», приводит к дальнейшему уменьшению степени неопределенности (энтропии) одной буквы балкарского текста. Поэтому при передаче такого текста

по линии связи можно еще уменьшить среднее число элементарных сигналов, затрачиваемых на передачу одной буквы. Для этого надо лишь подсчитать условную энтропию $H_2 = H_{\alpha_1}(\alpha_2)$ опыта α_2, состоящего в определении одной буквы балкарского текста, при условии, что нам известен исход опыта α_1, состоящего в определении предшествующей буквы того же текста (заметим, что при приеме очередной буквы сообщения мы всегда знаем уже предшествующую букву). Условная энтропия H_2 определяется следующей формулой:

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

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

В результате подсчета этих величин с помощью программы были получены следующие результаты:

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

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

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

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

. . . . . ,

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

Зная величину H_2, можно провести эксперимент и получить следующий результат:

лататодайы болалыдып къармни аханинген келауаннды ай шчы деры.

Зная величину H_3, можно провести эксперимент и получить следующий результат:

дермекди мемюрегенг бол хшюн ал жону да бла салдюрлик.

По звучанию эта «фраза» заметно ближе к балкарскому языку, чем фраза выписанная в первом случае и во втором случае.

Для H_5 моделирование привело к следующему результату:

жетгендиле бошады да да халкъ берсенг да къоюп келген.

Обсуждение результатов исследования

Среднее число элементарных сигналов, необходимое для передачи одной буквы текста, не может быть меньшим $\frac{H_\infty}{\log m}$; с другой стороны, возможно кодирование, при котором это среднее число сколь угодно близко к величине $\frac{H_\infty}{\log m}$. Разность $1 - \frac{H_\infty}{H_0}$, показывающую, насколько меньше единицы отношение «предельной энтропии» H_∞ к величине $H_0 = \log n$, характеризующей наибольшую информацию, которая может содержаться в одной букве алфавита с данным числом букв, Шеннон назвал избыточностью языка. В нашем случае имеем следующий результат:

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

Такая избыточность языка позволяет сокращать телеграфный текст за счет отбрасывания некоторых легко отгадываемых слов (предлогов и союзов); она же позволит легко восстановить истинный текст даже при наличии значительного числа ошибок в телеграмме или описок в книге.

Избыточность R является весьма важной статистической характеристикой языка. Для сравнения результатов, полученных для балкарского языка, приведем значения энтропии некоторых европейских языков:

Таблица 2

язык	Англ.	Немецк.	Франц.	Испанск.	Балк.
H_1	4,03	4,10	3,96	3,98	4,17

Для английского языка Шеннон получил следующие значения энтропий:

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

Для балкарского языка мы получили следующие результаты:

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

Опыты Шеннона [2, с. 669] показали, что величина H_{100}, по-видимому, заключена между 0,6 и 1,3бит. И для английского языка избыточность составляет порядка 80%. Для немецкого языка К. Кюнфмюллером [3, с. 265-272] было получено значение - 70%. Для французского языка избыточность была подсчитана Н.В. Петровой [4, с. 130-152] и она составила порядка 71%.

Список литературы

[1] Яглом А. М., Яглом И. М., *Вероятность и информация*, Наука, М., 1973, 512 с.

[2] Шеннон К., *Работы по теории информации и кибернетике*, ИЛ, М., 1963, 830 с.

[3] Кюпфмюллер К., “Энтропия немецкого языка”, *FTZ*, 1954, № 6, 265 – 272.

[4] Петрова Н. В., “Энтропия французского печатного текста”, *Известия Академии наук СССР. Серия литературы и языка*, **24**:1 (1965), 63–67.

Поступила в редакцию / Original article submitted: 29.03.2016