Modern Linguistics 现代语言学, 2026, 14(4), 111-118
Published Online April 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.144273

Hans汉斯

基于小型语料库的维吾尔语信息熵估算

高　昆

新疆大学中国语言文学学院，新疆　乌鲁木齐

收稿日期：2026年3月7日；录用日期：2026年3月22日；发布日期：2026年4月7日

摘　要

信息熵是一种衡量信息量的指标，在信息论中用于度量随机变量的不确定性。语言的信息熵是数学方法和语言学的结合，反映语言中每个字符的平均信息量，可以帮助我们了解语言中某一字符表达能力。本文以维吾尔语为研究对象开展维吾尔语信息熵估算研究。在简要论述香农三大定理与自然语言处理关系基础上，基于120万词的维吾尔语单语语料库开展频率统计，运用信息熵计算方法，将统计结果代入香农信息熵公式初步估算出了维吾尔语的零阶熵，并将估算结果与一些表音文字系统语言进行了对比。

关键词

维吾尔语，信息熵，香农三大定理，语料库

Estimation of Information Entropy in Uyghur Based on Small-Scale Corpora

Kun Gao

College of Chinese Language and Literature, Xinjiang University, Urumqi Xinjiang

Received: March 7, 2026; accepted: March 22, 2026; published: April 7, 2026

Abstract

Information entropy is a metric for measuring the amount of information, used in information theory to quantify the uncertainty of random variables. Linguistic information entropy combines mathematical methods and linguistics, reflecting the average information content of each character in a language, which helps us understand the expressive power of a particular character. This study focuses on Uyghur as the research subject, conducting an investigation into the estimation of Uyghur information entropy. After briefly discussing the relationship between Shannon's three theorems and natural language processing, the study performs frequency statistics based on a monolingual corpus of 1.2 million Uyghur words. Using information entropy calculation methods, the statistical

文章引用: 高昆. 基于小型语料库的维吾尔语信息熵估算[J]. 现代语言学, 2026, 14(4): 111-118.
DOI: 10.12677/ml.2026.144273

results are substituted into Shannon's entropy formula to preliminarily estimate the zero-order entropy of Uyghur. The estimated results are then compared with some languages of the abugida writing system.

Keywords

Uyghur Language, Information Entropy, Shannon's Three Major Theorems, Corpus

1. 引言

随着社会信息化程度的不断提高，自然语言处理(NLP)或为语言研究的一个重点与热点。而信息熵的计算就是进行自然语言处理的一项重要基础研究。教育部、国家语委、中央网信办明确提出，要以数字中文建设服务数字中国战略，全面加强语言文字信息化、规范化、标准化建设，以语言文字数字化与数据中文化深度融合，推动教育现代化和语言文字事业高质量发展。本项目充分发挥计算语言学作为“新文科”研究前沿的理论优势，灵活运用计算语言学研究方法估算维吾尔语信息熵。在理论层面具有以下几方面意义：一是，运用交叉学科研究方法开展具体语言问题研究，符合语言学研究发展趋势。二是，目前关于少数民族语言信息熵的研究较少，本项目可以一定程度上弥补少数民族语言在信息熵方面研究的空缺。

2. 各国对语言信息熵研究历程

信息熵是信息论的基本概念，也是信息论的基础。其描述所获取的信息中各可能事件发生的概率。从语言学角度来看，信息熵的基本作用是消除或减少人们对事件的不确定性，对于越模糊的事情，需要把它弄清楚所需要的信息量也就越多，其信息熵也就越大。比如：对于一道选择题，当做题者被告知正确答案为 C 后，那么他对问题的不确定性便被消除。此时，做题者从他所获取到的信息中获得了一定的信息量。通俗的讲，想要消除更多的不确定性，就需要更多的信息。因此，我们可以用在做题者接收到信息之前，其对题目答案不确定性程度的大小来表示语言符号所负荷的信息量。

1951 年，美国数学家香农首次运用数学方法测出了英语中包含在一个字母中的熵，并以此为基础提出了香农三大定律，奠定了现代信息科学基础。此后,随着科学的发展，在信息化时代需求下，人们依据香农三大定律，运用信息熵测量方法陆续测出了一些其他语言的信息熵。如：英语 4.03 比特；法语 3.98 比特；德语 4.10 比特；西班牙语 4.01 比特；俄语 4.35 比特；罗马尼亚语 4.12 比特等。至此，信息熵成为了语言文字信息化的一项基础数据，并广泛应用于信息编码、文字输入法以及文本自动处理等自然语言处理领域。

在国内，上世纪 70 年代末冯志伟先生通过对汉字手动查频，建立了 6 个不同容量的汉字频度表，采用逐渐扩大汉字容量的方法。首次计算出了汉字的零阶熵值是 9.65 比特。最后得出结论：当汉字容量不大时，汉字的信息熵随着汉字容量的增加而增加，当汉字容量达到 12,366 个字时，汉字的信息熵就不再增加，并将研究结果于 1984 年发表在《语文建设》中[1]。1995 年，冯志伟又根据英汉双语语料库的对比研究，进一步测定了在充分联系上下文的基础上包含在一个汉字中的熵，这个熵也就是汉字的“极限熵”。他测得，汉字的极限熵的平均值为 4.0462 比特[2]。尽管冯志伟先生认为他所计算的信息熵只是一

种猜测，但计算机发展落后的当时，通过对数量庞大的汉字进行手动查频计算出其零阶熵，也是一件极其困难的工作[3]。冯志伟先生的这项工作为汉字电脑端显示以及中文输入法制作奠定了语言学基础，也为汉字双字节编码奠定了理论基础。随后，随着信息科学理论发展与计算机计算能力的提升，语言信息熵的测算不断精确。80 年代末期，刘源运用汉字频度统计的方法，计算出汉字的零阶熵是 9.71 比特[4]。在 1999 年，黄萱菁等人在大规模语料的基础上求得的汉语的零阶熵为 9.62 比特、一阶熵为 6.18 比特、二阶熵为 4.89 比特[5]。孙帆等人用基于词的语言模型估计方法估算出汉语的极限熵值为 5.31 比特[6]。

关于我国少数民族语言文字信息熵的估算，由于当时计算机可读文本较少，所以起步较晚。在 2007 年第七届中文信息处理国际会议上，那日松、淑琴发表文章《蒙古文信息熵和拉丁转写研究》并估算出蒙古文名义字符的估算熵为 4.165 比特[7]。在语言文字信息熵计算方面大都采用基于语料库的统计方法计算信息熵。江荻以 20 余万字的藏文单语语料为基础，采用统计的方法，估算出了藏文字符的一阶熵其值为 3.9913 比特[8]。在 2020 年，完么扎西等人以 300 多万字的藏语单语语料为基础，采用概率统计的方法，估算了藏文字符的信息熵为 4.17 比特和藏文字的信息熵为 8.21 比特[9]。严海林等人同样采用概率统计的方法，以 4 千万字符的藏语单语语料为基础估计出了藏文“字丁”的零阶熵、一阶熵、二阶熵和三阶熵，以及冗余度，其值分别为 9.59 比特、4.80 比特、3.12 比特、2.70 比特和 0.72 比特[10]。对于少数民族语言文字信息熵的研究主要集中在蒙古文和藏文领域，对于维吾尔语信息熵估算方面的研究几乎为空白。

近些年随着科学的发展，计算机逐渐普及，各民族语言信息化文本不断丰富，少数民族语语料收集变得更加方便，这也为开展少数民族语言信息熵估算工作提供了便利，也使本研究开展具备了可能性。

3. 信息熵基本概念及其在语言学方面应用

3.1. 信息熵

20 世纪 40 年代末，香农(C.E. Shannon)借鉴了热力学中熵的概念，把信息中排除了无效信息后的平均信息量称为“信息熵”，同时也给出了描述信息熵的数学公式。信息熵的提出解决了对信息的量化度量问题。

信息熵是反映语言的数学面貌的一个重要的信息论参数，与热力学中的熵相似。他在著作 *The Bell System Technica Journal* 中提出，在信息论中的熵是信息不确定性的度量单位，也就是信息量的度量单位。他用公式 $H=-\log 2P$ 来表示信息量[11]。

信息熵的定义：如果一个符号集合 X，其中每个符号出现的概率分别为 $p_1, p_2, \cdots, p_n$，则其的熵为：

$$H(X)=-\sum_{i=1}^{n} p_i \log 2 p_i$$

信息论中采用比特(bit)作为信息量的单位[12]。

3.2. 香农三大定理与自然语言处理

香农三大定理是信息论的基础理论，是存在性定理，虽然在定理中并没有提供具体的信息编码实现方法，但却为通信信息理论的研究指明了方向。香农三大定理结合起来就构成了现代信息论的基础理论，三大理论之间相辅相成，相互联系，推动了通信技术的发展和应用，为现代通信数字理论的发展做出了巨大的贡献[13]。同样，在自然语言处理领域中，香农三大定理也有着很高的地位。

定理一：可变长无失真信源编码定理，假设有一个信息源，其中的信息都是来自 X 的 n 个字符。如果信息中的每一个字符都符合 $p(x)$ 分布，那么：

$$H(X) \le L_n < H(X)+\frac{1}{n}$$

其中 L_n 为所输入字符期望编码长度，因此，通过使用足够大的分组长度，可以获得一个编码，可以使其每字符期望码长任意地接近熵。可变长无失真信源编码定理是采用无失真最佳信源编码在失真度限制下，可以用较短的编码长度来表示信息，而不会引起信息传输的失真。它给出了在无损情况下，数据压缩的临界值。同时它也可以有效的提高通信系统的传输效率和质量，降低通信成本和复杂度。在自然语言处理中，我们可以将一个语言模型看作一个信息源，其中每个词汇都有一个概率分布，表示该词汇在该语境中出现的概率。通过计算语言模型的信息熵，我们可以评估其预测能力和模型的复杂度。可用于语言模型的比较与评估。

定理二：有噪信道编码定理，是关于编码存在的定理。只要信息传输速率小于信道容量，就存在一类编码，使得信息传输的错误概率可以任意小。它可以帮助我们设计更高效、更可靠的通信系统。在自然语言处理中，可以将该定理应用于语音识别和文本翻译等有干扰的任务中，通过优化代码的方式来提高模型的精确度和可靠性。

定理三：保真度准则下的信源编码定理，或称有损信源编码定理。在信息论中，保真度准则是衡量信息传输效果的一个重要指标。它可以用于评估信息传输过程中信息的失真程度，从而进一步优化信源编码方案。根据保真度准则，一个信源编码的失真率被定义为该编码对于每个信息符号所引入的平均失真程度。在这个定义下，如果一个编码方案的失真率小于预定的阈值，那么这个编码方案就可以被认为是有效的。在自然语言处理中，可以将该定理应用于文本压缩和数据压缩的任务中，通过压缩技术来减小储存空间的占用，提高传输效率。

根据字符编码的观点，语言符号的熵可视为该语言字符编码的平均最小码长。通过计算某语言字符的信息熵，可以确定该语言符号系统的较短编码长度，从而实现最佳的数据存储、管理和传输效率，以最小的成本和消耗为代价。例如，我们在处理一个有 10,000 个字符的英文文本，如果采用 1 个字节 8 比特代表一个符号的 ASCII 码进行编码，那么我们就需要 10000*8 = 80000 比特，来储存这个文本。如果我们统计出该文本每个字符出现频率所对应的概率，根据概率高低设计不同的编码长度，概率越高，编码长度越短。这样我们就可以把数据量压缩下去远远小于 80,000 比特，而香农提出的三大定理就是告诉我们如何计算数据压缩的极限。

香农的三大定理在自然语言处理中有着广泛的应用。它们可以用于语言模型的评估和比较、文本信息量的评估、语音信号的压缩和表示等方面。例如，在自然语言生成中，我们可以使用可变长无失真信源编码定理来评估生成模型的预测能力和复杂度，从而选择最优的模型。在语音识别中，我们可以使用有噪信道编码定理来评估不同的语音信号传输方式的可靠性和效率，从而选择最优的传输方式。在文本分类中，我们可以使用保真度准则下的信源编码定理来对文本进行压缩和表示，从而提高分类效果和减少计算成本。

此外，香农三大定律也为自然语言处理研究提供了重要的理论基础。它们揭示了信息传输和处理中的基本规律和限制，为自然语言处理技术的设计和优化提供了指导和启示。

4. 维吾尔语语料库的建设及语料处理

4.1. 维吾尔语语料库构建

随着计算机的迅猛发展，大量计算机可读文本不断产生，电脑语料库容量大，资料真实，信息提取准确，因此利用语料库对语言进行研究已经发展成为一个跨世纪跨学科的语言研究学科。

语料库是一种被广泛用于自然语言处理和语言学研究的数据集合。它是由收集、整理、标注和储存大量的自然语言文本构成的，并可用于研究自然语言的语法、语义、语用、文本分类、词汇分析、机器翻

译、信息检索等领域。语料库的规模可以从小到大，可以包含数千个语言样本，也可以包含数亿个语言样本。语料库可以通过手工构建，也可以是从网络、出版物、语音录音等不同渠道自动采集。本文所使用的语料来源于新疆大学 2018~2019 级硕士毕业生翻译材料，共计 120 万词，材料中的所有作品均由维吾尔族母语作者撰写，且均已正式出版，确保了语料的语言真实性和规范性。

4.2. 语料处理

维吾尔语是一种属于阿尔泰语系的语言，它的文字系统基于阿拉伯字母。维吾尔语的书写方式是从右向左、从上到下。维吾尔语共有 32 个字母，其中，分为 24 个辅音字母和 8 个元音字母。在书写维吾尔语时，根据字母出现在词语中位置的不同，这 32 个字母共有 126 种不同的书写形式可供使用。用空格将单词和单词分开，使其更易于阅读和理解。

在计算机处理维吾尔语老文字时，由于维吾尔文采用基于阿拉伯字母的书写系统，其 Unicode 编码中，出现在词首的元音通常由两个独立字符组合表示(例如，元音[ɑ]在词首写作“ئا”，由“ئ”和“ا”两个字符组成)，这导致计算机统计时会将一个音素误计为两个字符，从而扭曲后续的概率分布。为了准确反映维吾尔语的语音单位，我们采用国际音标对语料进行转写。国际音标是一种记录语言中的音素的符号，它是一种公认的书写符号，一个音素只表示一个音标，一个音标只代表一个音素。国际音标是为了方便记录语言中的音素而制定的，每个符号都有特定的含义和用途[14]。这些符号的表示方式是基于其在语言学中的定义和使用方式，以确保它们能够准确地表示语言中的音素，因此可有效消除因计算机编码带来的计数偏差。此外，转写过程中还去除了语料中的标点符号、数字及特殊字符，仅保留纯音素序列，以确保统计对象为语言的核心语音单元。

将语料划分为一百份文本文件，并导入至 Python 脚本中。为准确实现字符转换，脚本利用 Python 字典存储维吾尔语老文字字符与国际音标之间的映射关系；由于字典要求每个键唯一对应一个值，这与国际音标“一音一符”的原则高度契合。随后，脚本对语料进行批量处理：将维吾尔语老文字转换为国际音标，同时过滤掉所有标点符号、数字及特殊字符。

5. 维吾尔语信息熵的估算

5.1. 维吾尔语信息熵估算公式

计算语言的熵是使用数学方法和语言文字结合，语言的熵反映了语言中每个字符所携带的平均信息量。在测定语言的熵值时，我们通常只考虑语言符号出现概率的不同，而不必考虑它们出现概率之间是否相互影响。因此，用传统的信息论公式计算得出的数值是静态平均信息熵。根据信息论的基本原理，这个熵也可以被称为“零阶熵”。本文中估算的维吾尔语信息熵为维吾尔语的零阶熵。

以维吾尔语为例，我们将维吾尔语看作一个离散信源。若 X (维吾尔语中的维文字符或单词)为离散随机变量，则变量 X 元素的集合及其每个元素出现的概率 Pi 分别为：

$$X=\{x_1,x_2,\cdots,x_n\}$$

$$P_i=P[X=x_i]$$

由此可得：

$$\begin{bmatrix} X \\ p(x) \end{bmatrix}=\begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_n \\ p_1 & p_2 & \cdots & p_i & \cdots & p_n \end{bmatrix} \text{且} \sum_{i=1}^{n} p_i=1$$

其中，P_i 是离散信源各元素发生的概率。若离散信源中每个元素的出现与上下文无关，且每个元素出现的概率相等，则该离散信源含有的信息量(H_0)为：

$$H_0 = \log 2\frac{1}{p_i} = -\log 2 p_i$$

但是，在维吾尔语语言文字中，构成维吾尔语句子的各元素 X_i 出现的概率是不可能相等的。若暂时不考虑各个元素的上下文相关性，则该离散信源中各个元素 X_i 含有的平均信息量(H)为：

$$H = \sum_{i=1}^{n} p_i \log 2\frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log 2 p_i$$

在信息论中，H 被定义为信息熵。容易证明 $H < H_0$，由此我们就可以通过公式计算出维吾尔语字符中所包含的平均信息量，即忽略上下文关系的零阶熵。

5.2. 维吾尔语信息熵估算过程

语言通常都是由一组符号的集合构成的信息源，这些符号可以代表不同的音节、单词或语素。不同的语言文字都有不同的书写方式，如：汉字是汉语的书写符号，英语和维吾尔语采用不同的字母作为书写符号。字母表则是一个由字母组成的符号库，用来表示一定的发音规则和语言含义，而且这些字母在各种文本中的出现也有一定的规律性。因此本文将采用基于字符统计的信息熵计算方法，估算维吾尔语的信息熵。首先，统计各个字母在语料库中出现的次数，之后，计算字母的频率分布，最后，应用香农的信息熵公式估算出维吾尔语的信息熵。

根据上述过程，首先统计各个字母在语料库中出现的次数，将语料导入至 word 文件中并对语料中每个字母出现的次数进行手动查频，为了防止人工手动查频出现错误，因此又用代码对语料进行自动查频，对照两次查频数据，得到总字符数为 7,442,799 个。

第二步：获取总字符数下各字母概率分布，根据上一节中介绍的相关方法，对语料库的的总字符数与各字母在语料库中出现的次数进行统计，并根据字母出现频率计算各字母的概率分布：

$$P_i = \frac{\text{该字母出现频次}}{\text{总字母数}} (i = 1 \sim 32)$$

得到每个字母出现概率，见表 1。

Table 1. Uyghur language frequency and probability statistics table
表 1. 维吾尔语频次、概率统计表

老文字	国际音标	频次	概率(P_i)	老文字	国际音标	频次	概率(P_i)
ئا	ɑ	691,849	0.092955486	ق	q	273,878	0.036797715
ئە	ɛ	502,172	0.067470853	ك	k	231,435	0.031095157
ب	b	188,428	0.02531682	گ	g	79,984	0.010746495
پ	p	176,682	0.02373865	ڭ	ŋ	112,910	0.015170368
ت	t	326,612	0.043882953	ل	l	451,550	0.06066938
ج	ʤ	25,877	0.003476783	م	m	291,935	0.039223819
چ	ʧ	101,780	0.013674963	ن	n	476,422	0.064011133
خ	χ	47,804	0.006422852	ھ	h	62,563	0.008405843
د	d	286,155	0.038447229	ئو	o	142,177	0.019102625

续表

ر	r	348,079	0.046767218	ئۇ	u	302,398	0.040629607
ز	z	112,976	0.015179236	ئۆ	ø	64,101	0.008612486
ژ	ʒ	389	5.22653E−05	ئۈ	y	107,093	0.014388807
س	s	171,361	0.023023731	ۋ	w	59,874	0.008044554
ش	ʃ	150,110	0.020168488	ئې	e	145,327	0.019525853
غ	ʁ	113,776	0.015286722	ئى	i	1173,799	0.157709351
ف	f	3049	0.000409658	ي	j	220,254	0.029592899

第三步：估算信息熵，根据香农的信息熵计算公式，估算维吾尔语信息熵。即将各个字母出现的概率带入信息熵公式计算：

$$H(X)=-\sum_{i=1}^{n=32}p_i\log 2p_i$$

用 excel 表格对数据进行计算得到：比特

$$\begin{aligned}H(X)&=-(0.0929\times\log_2 0.0929+\cdots+0.0295\times\log_2 0.0295)\\&\approx-(-4.409678)\\&\approx 4.431301(\text{比特})\end{aligned}$$

5.3. 维吾尔语信息熵估算结果

通过上述计算，我们得到维吾尔语字母的零阶熵为 4.431301 比特。为验证结果的合理性，我们将其与同为拼音文字的俄语(零阶熵 4.35 比特)进行比较。需要注意的是，俄语的熵值是基于类似统计方法得到的，但其语料库规模、文本体裁及语言结构存在差异。俄语使用 33 个西里尔字母，而维吾尔语虽仅有 32 个基础字母，但实际书写中字母存在词首、词中、词末等变体形式，本研究通过国际音标转写统一为音素层面，因此统计单位与俄语字母相当。此外，俄语语料通常涵盖文学、新闻等多领域，而本文语料主要来自硕士翻译材料，可能偏向书面语，这可能导致频率分布略有不同，但总体而言，相近的字母数量和熵值表明维吾尔语在字母层面的信息负载与俄语等语言处于同一量级。

从字母频率分布看，音素/i/、/ɑ/、/ɛ/、/l/、/t/、/r/、/n/的出现频率最高，累计占比达 53%。这一分布特性对实际应用具有重要指导意义。例如，在设计维吾尔语最优编码方案时，可依据频率采用变长编码(如哈夫曼编码)，为高频字母分配较短码字，从而降低平均码长。理论上，最佳编码的平均码长应接近信息熵值。若以当前频率分布构建哈夫曼编码，其平均码长可计算并与 4.431 比特比较，以评估编码效率。类似地，在维吾尔文输入法设计中，可将高频音素映射到便于触及的键位，或通过词频预测减少击键次数，从而提升输入效率。

需要指出的是，本研究所估算的零阶熵仅为初步结果，存在一定的不确定性。首先，语料规模为 120 万词，虽能反映基本频率特征，但受限于样本量，低频字母(如ʒ，频次仅 389)的频次极低，其概率估计的抽样误差较大。根据信息熵的方差近似公式，可粗略估计熵值的标准误约为 0.02 比特量级，对应的 95% 置信区间约为[4.39, 4.47]比特。然而，这一区间仅为理论近似，实际不确定性需通过自助法等重采样技术进一步精确评估。其次，语料来源相对单一，未充分覆盖口语、科技、新闻等多种文体，可能导致频率分布存在偏差。

6. 结论

本文引用信息论的方法研究了维吾尔语文字的信息熵，并在 120 万词语料的基础上，采用字符统计的方法对维吾尔语的信息熵进行了估算，所求得的零阶熵约为 4.431301 比特。得到的结果已经相当接近其他拼音文字的信息熵。这个研究成果可以应用在计算机维吾尔语信息处理的研究领域，如：建立语言模型，文本压缩，机器翻译，评估维吾尔语生成式语言模型质量等。

在刚才我们估算维吾尔语字母的信息熵时，只考虑了每个字母在文本中出现的概率，而没有考虑这些字母之间的相互影响。因此，我们得到的只是维吾尔语的零阶熵。实际上，语言符号的出现概率是相关的，彼此相互影响的。比如，在维吾尔语词汇中一个音节中只有一个元音字母，所以在同一个音节中，当已经出现了一个元音字母时，下一个字母是元音的概率为零。在充分考虑上下文关联后计算得出的信息熵被称为极限熵。基于极限熵的数值，我们可以深入研究如何提高字符传输速度的编码方法。这种方法不仅针对单个字符，还针对整个文本进行编码，从而使平均码长缩短，更接近极限熵。

需要说明的是，本研究存在一定的局限性。首先，120 万词的语料规模虽能提供基础频率统计，但对于低频字母的概率估计仍不够稳定，且语料来源偏重书面翻译文本，未能涵盖口语、科技、新闻等多领域，可能导致频率分布存在系统性偏差。未来若采用更大规模、更均衡的语料库，有望将熵值估计精度提升至±0.01 比特以内。其次，本研究仅估算了零阶熵，即假设字母独立同分布，而实际语言中字母之间存在强烈的上下文关联，因此零阶熵仅反映了信息量的上限。后续研究可进一步计算一阶、二阶乃至高阶熵，并最终逼近极限熵，从而为维吾尔语文本压缩、语言模型评估等应用提供更精确的理论依据。尽管存在上述局限，本研究作为维吾尔语信息熵的首次估算，仍为相关领域提供了基础数据和方法参考。

参考文献

[1] 冯志伟. 汉字的熵[J]. 文字改革, 1984(4): 12-17.

[2] 冯志伟. 汉字的极限熵[J]. 中文信息, 1996(2): 53-56.

[3] 关于汉字的熵和极限熵致编辑部的一封信[J]. 中文信息学报, 1998(1): 64-65.

[4] 刘源. 汉语字词的概率分布, 熵及冗余度[C]//中文信息处理国际会议论文集. 1987.

[5] 黄萱菁, 吴立德, 郭以昆, 刘秉伟. 现代汉语熵的计算及语言模型中稀疏事件的概率估计[J]. 电子学报, 2000(8): 110-112.

[6] 孙帆, 孙茂松. 基于统计的汉字极限熵估测[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论集. 北京: 清华大学出版社, 2006: 550-559.

[7] 那日松, 淑琴. 蒙古文信息熵和拉丁转写研究[C]//中国计算技术与语言问题研究——第七届中文信息处理国际会议论集. 北京: 电子工业出版社, 2007: 793-796.

[8] 江荻. 藏语文本信息处理的历程与进展[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集. 北京: 清华大学出版社, 2006: 91-105.

[9] 完么扎西. 现代藏语信息熵的估算及语言模型的复杂度[J]. 电子技术与软件工程, 2020(17): 213-215.

[10] 严海林, 江荻. 藏文大藏经信息熵研究[C]//那顺乌日图, 陈玉忠. 中国少数民族多文种信息处理研究与进展. 2004: 1-6.

[11] Shannon, C.E. and Weaver, W. (1949) The Mathematical Theory of Communication. The University of Illinois Press.

[12] Shannon, C.E. (1951) Prediction and Entropy of Printed English. *Bell System Technical Journal*, **30**, 50-64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

[13] 邓烨. 从通信系统的收发联合优化看香农三大定理的内在联系[J]. 中国新通信, 2012, 14(6): 78-80.

[14] 帕丽旦·木合塔尔, 热依曼·吐尔逊, 吾守尔·斯拉木, 买买提阿依甫. 维吾尔文本转换国际音标系统设计与实现[J]. 信息通信, 2017(5): 97-99.