A Convergent Gambling Estimate of the Entropy of English

THOMAS M. COVER, FELLOW, IEEE, AND ROGER C. KING, STUDENT MEMBER, IEEE

Abstract—**In his original paper on the subject, Shannon found upper and lower bounds for the entropy of printed English based on the number of trials required for a subject to guess subsequent symbols in a given text. The guessing approach precludes asymptotic consistency of either the upper or lower bounds except for degenerate ergodic processes. Shannon's technique of guessing the next symbol is altered by having the subject place sequential bets on the next symbol of text. If S_n denotes the subject's capital after n bets at 27 for 1 odds, and if it is assumed that the subject knows the underlying probability distribution for the process X, then the entropy estimate is $\hat{H}_n(X)=(1-(1/n)\log_{27} S_n)\log_2 27$ bits/symbol. If the subject does not know the true probability distribution for the stochastic process, then $\hat{H}_n(X)$ is an asymptotic upper bound for the true entropy. If X is stationary, $E\hat{H}_n(X)\rightarrow H(X)$, $H(X)$ being the true entropy of the process. Moreover, if X is ergodic, then by the Shannon–McMillan–Breiman theorem $\hat{H}_n(X)\rightarrow H(X)$ with probability one. Preliminary indications are that English text has an entropy of approximately 1.3 bits/symbol, which agrees well with Shannon's estimate.**

1. Introduction

THE GOAL of this paper is to develop an accurate estimate of the entropy of printed English. For a discrete random variable Y, the entropy associated with Y is $H(Y)=-\sum_i p(y_i)\log_2 p(y_i)$ where Y takes the value y_i with probability $p(y_i)$. Let printed English be represented by the symbol X and consist of strings of the form $(\cdots,x_{-1},x_0,x_1,\cdots)$. If we assume English to be a stationary random process, as we shall in this paper, then we define the entropy $H(X)$ of the process X to be

$$H(X)=\lim_{n\rightarrow\infty} H(X_n|X_{n-1},\cdots,X_1)=\lim_{n\rightarrow\infty}\frac{1}{n}H(X_1,\cdots,X_n). \tag{1}$$

Alternative characterizations of $H(X)$ are

$$\begin{aligned} H(X)&=\lim_{n\rightarrow\infty} H(X_0|X_{-1},\cdots,X_{-n})\\ &=H(X_0|X_{-1},X_{-2},\cdots)\\ &=\lim_{n\rightarrow\infty}\frac{1}{n}H(X_n,\cdots,X_1|X_0,\cdots,X_{-k}), \end{aligned} \tag{2}$$

which follow using the boundedness and continuity of $h(p)=-p\log p-(1-p)\log(1-p)$. In addition, if English is an ergodic process, then the Shannon–McMillan–Breiman theorem states

$$-\frac{1}{n}\log_{27} p(X_1,\cdots,X_n)\rightarrow H(X) \quad \text{a.e.} \tag{3}$$

If printed English is indeed an ergodic process, then for sufficiently large n a good estimate of $H(X)$ can be obtained from knowledge of $p(\cdot)$ on a randomly drawn string $(X_1,\cdots,X_n)$.

An additional comment is in order concerning the meaning of the phrase "the entropy of English." It should be realized that English is generated by many sources, and each source has its own characteristic entropy. The operational meaning of entropy is clear. It is the minimum expected number of bits/symbol necessary for the characterization of the text. A gambling approach will yield an estimate of the entropy that is consistent with the above operational meaning whether or not the assumption of ergodicity for the stochastic process of English text is satisfied.

Just as the entropy rates associated with various authors differ, so there are different entropy estimates associated with different gamblers. The difference in the entropy estimates is associated with the amount of money that each of the gamblers can make on the sequence and is profoundly affected by the gambler's ability to accurately quantify his previous empirical experience with English. Thus an intelligent well-educated gambler will do better than a gambler untrained in quantitative thinking who is relatively unfamiliar with the language. Nonetheless, it will be true that there is an upper bound on how well a gambler can do. If there were no such bound, then the true entropy of the creative process of the writer would be zero and his writing totally predictable. This upper bound yields the entropy estimate we seek.

An extensive bibliography of papers relating directly to Shannon's paper [1] on the entropy of English is included. A brief discussion of these papers follows.

Several papers provide important theoretical material. Maixner [2] helps to clarify the details behind the derivation of Shannon's lower bound to Nth-order entropy approximations. Savchuk [3] gives necessary and sufficient conditions on the source distribution for Shannon's bounds to hold with equality. Background on entropy estimate limitations can be found in [4]–[8]. Important

Manuscript received November 1, 1976; revised November 4, 1976. This work was supported in part by the United States Air Force Office of Scientific Research under Contract F44620-74-C-0068, and in part by the National Science Foundation under Contract ENG 76-03684. This work was presented at the IEEE International Symposium on Information Theory, Ronneby, Sweden, 1976.

T. M. Cover is with the Departments of Electrical Engineering and Statistics, Stanford University, Stanford, CA.

R. C. King was with the Department of Electrical Engineering, Stanford University, Stanford, CA. He is now with SATCOM System Engineering, M.I.T. Lincoln Laboratory, Lexington, MA 02173.

0018-9448/78/0700-413$00.75

factors involved in eliciting probability assessments from experimental subjects can be found in Savage [9]. Consistent objective estimates of the entropy of finite alphabet ergodic processes with unknown distribution appear in Bailey [10]. Chomsky [11], Mandelbrot [12], Berry [13], Bell [14], and Yngve [15] all give important insight into the structure of language from the viewpoint of information theory.

A different estimation technique can be found in Newman and Gerstman [16]. This paper has been quoted extensively in psychology literature, but the theory does not include a proof of the consistency of its entropy estimate.[1]

Several papers entend or comment on Shannon's empirical results for English text. Grignetti [17] recalculates Shannon's estimate of the average entropy of words in English text. Burton and Licklider [18] use longer passages of text for Shannon's estimate. Paisley [19] studies entropy variations due to authorship, topic, structure, and time of composition. Treisman [20] comments on contextual constraints in language, and Miller and Coleman [21] provide more data on the entropy of English using Newman and Gerstman's technique.

Kolmogorov (as characterized in [64, p. 257] and [98, p. 160]) argues that the following strategy will consistently estimate conditional probabilities and hence the entropy. (See also Savage [9].) Suppose that the subject knows the conditional probability p_i of the event that the next symbol in the text is the ith letter of the alphabet. In each experiment the subject has to name these probabilities. Proceeding through the text, one calculates the running average of the logarithms $-\log p_k$, where k denotes the actual outcome of the experiment. If the p are correct, then this average converges to $H(\boldsymbol{X})$. We shall find that this analysis arises as a natural by-product of the gambling estimate treated here, thus providing an operational motivation for this estimate. Moreover, it will be shown that an incorrect assessment of the conditional p_i leads to an overestimate of $H(\boldsymbol{X})$.

Many other papers [22]–[31] apply varied techniques to estimating the entropy of different languages. Tzannes *et al.* [32] and Parks [33] both measure the entropy of digitized images.

Shannon's or related estimates are used in many wide-ranging applications in [34]–[63]. The psychology literature is particularly rich in entropy estimates.

An important reference work on the subject is the book by Yaglom and Yaglom [64], which contains an extensive bibliography. Translations are given of the entries [65]–[97] in this bibliography that bear directly on the problem at hand and may be of interest to future workers in the area. Another thorough reference work, also with an extensive bibliography, is the book by Weltner [98]. The references in [98] require no translation. Additional references [99]–[101] were suggested by the referees. The reader is advised that some references in the psychology literature and many references in [64], [98] have not been included.

[1]Moreover, the authors make frequent use of the quantity $D(n)=1-(H(X_n|X_1)/H(X_1))=I(X_1;X_n)/H(X_1)$ together with an assertion based on empirical evidence that $D(n)=1/n^2$. However, for aperiodic Markov chains of arbitrary order, it is easily proved that $D(n)=c\rho^n$, where $\rho<1$.

II. Shannon's Estimate

Shannon [1] found an upper bound to the entropy of printed English and a lower bound to the Nth order approximation of English by eliciting knowledge of $p(\cdot)$ from a subject through the use of a guessing scheme. A subject is shown $N-1$ consecutive symbols of unfamiliar text. He is then instructed to guess the next letter in the passage. Guesses are made in decreasing order of conditional probability until a correct guess occurs. Defining $\hat{q}_i^N$ to be the relative frequency of times the subject required i guesses to discover the correct letter given the $N-1$ previous letters, we can express Shannon's upper bound as

$$H(\boldsymbol{X}) \leqslant -\sum_{i=1}^{27} \hat{q}_i^N \log_2 \hat{q}_i^N \tag{4}$$

where n samples have been taken to establish $\hat{q}_i^N$ and blanks are included to give an alphabet of 27 symbols. We note that the upper bound is loose for three reasons: 1) N is finite, 2) $\hat{q}_i^N$ is determined by a mixture of $\hat{q}_i^N$ conditioned on the past, 3) the sample size n is finite, and thus $\hat{q}_i^N$ is a random variable that has not yet converged to its mean. The first two reasons cause the upper bound to be strictly greater than $H(\boldsymbol{X})$, and the third implies that the expectation of the upper bound will be strictly greater than the upper bound of the expectation. Shannon's bounds are derived for a subject who knows the true conditional probabilities $p(X_n|X_{n-1},\cdots,X_{n-N+1})$. For such a subject, Shannon defines q_i^N to be equal to the probability that the subject requires i guesses to discover the correct letter following a sequence of $N-1$ symbols. The basis for Shannon's empirical estimates are the following bounds:

$$\text{i)}\quad \sum_{i=1}^{27} i\left(q_i^N - q_{i+1}^N\right)\log i \leqslant H(X_n|X_{n-1},\cdots,X_{n-N+1}) \leqslant -\sum_{i=1}^{27} q_i^N \log q_i^N$$

$$\text{ii)}\quad H(\boldsymbol{X}) \leqslant H(X_n|X_{n-1},\cdots,X_{n-N+1}) \leqslant -\sum_{i=1}^{27} q_i^N \log q_i^N. \tag{5}$$

Thus we see that the lower bound (the first inequality in (5i)) is really a lower bound to an upper bound (the first inequality in (5ii)) and is therefore of limited meaning.

Define a map $\phi_N : \boldsymbol{X} \to \boldsymbol{S}_N$, where $\boldsymbol{S}_N$ is a new process taking values in $\{1,2,\cdots,27\}$. The map is determined by

$$\phi_N(X_n,X_{n-1},\cdots,X_{n-N+1}) = j, \quad \text{if } X_n \text{ is the } j\text{th most likely symbol, given } X_{n-1},\cdots,X_{n-N+1}.$$

Assuming X is an ergodic process, it is shown in Shannon [1] and Maixner [2] that

$$H(X_n|X_{n-1},\cdots,X_{n-N+1})=H(S_n|S_{n-1},\cdots,S_{n-N+1}). \tag{6}$$

The second bound above follows immediately, since

$$\begin{aligned} H(\boldsymbol{X}) &\leqslant H(X_n|X_{n-1},\cdots,X_{n-N+1}) \\ &= H(S_n|S_{n-1},\cdots,S_{n-N+1}) \\ &\leqslant H(S_n)=-\sum_{i=1}^{27} q_i^N \log q_i^N. \end{aligned} \tag{7}$$

The distribution over which the entropy is calculated to find the upper bound is a very rough approximation to the distribution including past information. The point is that *no* guessing game of this type can in general estimate $H(\boldsymbol{X})$ accurately if $H(S_n|S_{n-1},\cdots,S_{n-N+1})<H(S_n)$. A derivation of Shannon's lower bound, the first bound above, can be found in Shannon [1], Maixner [2], and Savchuk [3]. The upper and lower bounds are generally not equal, and the true entropy $H(\boldsymbol{X})$ generally falls strictly below the upper bound.

III. Gambling Approach

The essence of the gambling estimate lies in an optimal gambling scheme. Instead of guessing symbols and counting the number of guesses until correct as in Shannon's technique, the subject wagers a percentage of his current capital in proportion to the conditional probability of the next symbol in the alphabet conditioned on the past. This process is repeated on subsequent symbols of text with the subject accumulating S_n dollars after n wagers. If we have an ideal subject and he divides his capital on each bet according to the true probability distribution on the next symbol, we shall show in this section that, with probability one,

$$\left(1-\frac{1}{n}\log_{27} S_n\right)\log_2 27 \to H(\boldsymbol{X}) \quad \text{bits. a.e.} \tag{8}$$

This is an extension of the work of Kelly [102] and Breiman [103] on gambling on favorable independent trials to gambling on ergodic processes [104]. If the subject bets according to a distribution other than that of the process, then with probability one

$$\left(1-\overline{\lim_{n\to\infty}}\ \frac{1}{n}\log_{27} S_n\right)\log_2 27 \geqslant H(\boldsymbol{X}). \tag{9}$$

Let the English alphabet, augmented by blanks, be represented by X and denote the set of all finite strings of symbols from X by X^*.

Definition: Let $b(\cdot|\cdot): X\times X^*\to\mathbb{R}$ be called a *sequential gambling system* if the following conditions are satisfied:

i) $$b(\cdot|\cdot)\geqslant 0$$

ii) $$\sum_{x_{k+1}} b(x_{k+1}|x_k,\cdots,x_1)=1. \tag{10}$$

Definition: Associated with every gambling system is a *capital function* defined recursively by

$$S_0(\Lambda)=1 \qquad \text{(where } \Lambda \text{ is the null string)}$$

$$S_{n+1}(x_1,\cdots,x_{n+1})=27b(x_{n+1}|x_n,\cdots,x_1)S_n(x_1,\cdots,x_n), \quad n=1,2,\cdots. \tag{11}$$

Thus if sequential bets are placed on a sequence $x\in X^\infty$ and at time k a proportion $b(x_{k+1}|x_k,\cdots,x_1)$ of the current capital is bet on the outcome x_{k+1}, with fair odds being paid, the resultant capital is $S_{k+1}(x_1,\cdots,x_{k+1})$.

Definition: $S: X^*\to\mathbb{R}$ is *achievable* if there exists a sequential gambling scheme with initial capital $S(\Lambda)=1$ achieving $S(x)$ for all $x\in X^*$.

Theorem 1: The capital function $S: X^*\to\mathbb{R}$ is achievable by a sequential gambling scheme if and only if for all n and for all $x\in X^*$, $S_n(x_1,\cdots,x_n)27^{-n}=p(x_1,\cdots,x_n)$ are marginal distributions for some stochastic process $\{X_i\}_{i=1}^\infty$.

The proof is given in Cover [104].

Theorem 2: For any sequential gambling scheme b', $(n-E\log_{27} S_n(x_1,\cdots,x_n))\log_2 27\geqslant H(X_1,\cdots,X_n)$ for all n, with equality if and only if $b'=b^*$, where

$$b^*(x_{k+1}|x_k,\cdots,x_1)=p(x_{k+1}|x_k,\cdots,x_1), \quad k=1,2,3,\cdots. \tag{12}$$

The proof is given in Cover [104]. See Kelly [102] for the same result for i.i.d. processes.

Thus we have the intuitively satisfying result that to gamble optimally we simply place bets according to the conditional probability of possible outcomes given the past. Such a scheme is often called "proportional gambling."

Theorem 3: Let $\{X_n\}_{n=1}^\infty$ be an ergodic process with distribution p.

i) If the b^*-scheme is used, then

$$(1-(1/n)\log_{27} S_n)\log_2 27\to H(\boldsymbol{X}) \text{ a.e.}$$

ii) For any other scheme b,

$$\Pr\left\{\left(1-\overline{\lim}\ (1/n)\log_{27} S_n\right)\log_2 27\geqslant H(\boldsymbol{X})\right\}=1.$$

Proof of i):

$$\left(1-\frac{1}{n}\log_{27} S_n\right)\log_2 27=\left(1-\frac{1}{n}\log_{27} 27^n p(X_1,\cdots,X_n)\right)\cdot\log_2 27\to H(\boldsymbol{X}) \quad \text{a.e.} \tag{13}$$

by the SMB theorem.

Proof of ii): The proof uses the AEP. See [104].

Using (2) we can easily extend the b^*-scheme to a dependence on the past. The proof of the following theorem is straightforward given our previous statements.

Theorem 4: If $\boldsymbol{X}$ is an ergodic process and a gambling scheme $\bar{b}$ defined by

$$\bar{b}(x_n|x_{n-1},\cdots,x_{-k})=p(x_n|x_{n-1},\cdots,x_{-k}) \quad (14)$$

is used, then the induced capital function S_n obeys

$$\left(1-\frac{1}{n}\log_{27} S_n\right)\log_2 27\to H(\boldsymbol{X}) \quad \text{a.e.} \quad (15)$$

The $\bar{b}$ gambling scheme provides the tool with which to find an asymptotically correct estimate of the entropy of printed English. The subject inspects the text thoroughly up to a point x_0. Starting with $S(\Lambda)=1$ unit at time zero, the subject places bets according to the $\bar{b}$ scheme on the next outcome x_1. Fair odds are paid (27 for 1), and the process continues to symbol x_n of the text at which time the subject has S_n dollars where

$$S_n=S_n(x_1,\cdots,x_n)=\bar{b}(x_1,\cdots,x_n|x_0,\cdots,x_{-k})27^n. \quad (16)$$

By Theorem 4,

$$\left(1-\frac{1}{n}\log_{27} S_n\right)\log_2 27\to H(\boldsymbol{X}) \quad \text{bits/symbol a.e.} \quad (17)$$

We use the $\bar{b}$-scheme, because letting the subject inspect as much past text as he wishes allows him to formulate the best subjective opinion he can of the true statistical distribution of the given text. Roughly speaking, convergence to the entropy of the process should take place faster than if the past were limited.

IV. Education of the Gambler

How is it that asking a subject to gamble will elicit an accurate entropy estimate? We have already argued that there is no way to gamble in such a manner that the expected log capital $E\log S_n$ exceeds $n-H(X_1,\cdots,X_n)$, but how can we be assured that ordinary human gamblers will choose to achieve this limit?

First let us observe that each gambler has the vague motivation to increase his capital to a large amount with high probability. We present the gambler with three arguments for the proportional gambling scheme.

1) Maximizing the expected logarithm of the return is achieved by proportional gambling. Thus if the gambler's utility function is logarithmic in money, betting in proportion to the probabilities is optimal. Of course, we do not believe that a given gambler's utility function is precisely logarithmic in money, so this point is not emphasized.

2) The results of Kelly [102] and Breiman [103] indicate for independent gambles that maximizing the expected logarithm of the return on each gamble (which is achieved by proportional gambling) will cause one's money to grow to infinity at the highest possible rate on the condition that one does not go broke. We then argue (as shown in Cover [104]) that if the stochastic process is ergodic then conditional proportional gambling will cause S_n to grow to infinity at the highest possible rate, with probability one. Moreover, we show that even if one is willing to go broke with probability $\lambda>0$, conditional proportional gambling is still optimal and the growth rate of capital is unchanged, i.e., independent of λ for $0\leqslant\lambda<1$. The proof is similar to the strengthening of Shannon's weak converse to Wolfowitz's strong converse.

3) We can show that proportional gambling is also competitively best. This is exciting because it is consistent with the motivation of many gamblers approached for this project, in the sense that they were interested in achieving more money on the given sequence than any of the other participants. Let $b(x)$ be any gambling scheme on the random variable X, $P(X=x)=p(x)$, $x\in X$. Thus $\sum b(x)=1$, $b(x)\geqslant 0$. Let $O(x)$ be the odds offered given that alternative x is the outcome of the drawing of the random variable X. Thus the gambling scheme b induces the capital $S(x)=O(x)b(x)$, with probability $p(x)$. Consider the proportional gambling scheme $b^*(x)=p(x)$, with induced capital $S^*(x)=O(x)p(x)$ with probability $p(x)$. Then we have the result [104].

Theorem 5:

$$P\{S(X)\geqslant tS^*(X)\}\leqslant 1/t, \qquad \text{for } t\geqslant 0. \quad (18)$$

Proof:

$$\begin{aligned}P\{S(X)\geqslant tS^*(X)\}&=p\{b(X)O(X)\geqslant tp(X)O(X)\}\\&=P\{p(X)\leqslant b(X)/t\}\\&=\sum_{x:p(x)\leqslant b(x)/t} p(x)\leqslant\sum b(x)/t\\&=1/t.\end{aligned} \quad (19)$$

Corollary : Let $\{X_i\}_{i=1}^{\infty}$ be a stochastic process. Let $b^*(x_{k+1}|x_1,\cdots,x_k)=p(x_{k+1}|x_1,\cdots,x_k)$ and let $b(\cdot)$ be any other sequential gambling scheme. Then

$$P\left(\frac{1}{n}\log_2 S(X_1,\cdots,X_n)\geqslant\frac{1}{n}\log_2 S^*(X_1,\cdots,X_n)+\frac{t}{n}\right)\leqslant 2^{-t}, \text{ for all } t. \quad (20)$$

Summarizing, we see that proportional gambling is best for logarithmic utility functions, is competitively best, and causes one's capital to grow at the highest possible interest rate. Thus it behooves a gambler motivated by any of these three considerations to gamble in a proportional manner, alloting his next bet independently of the odds but according to the conditional probability distribution of the next symbol given the available past.

V. Operational Meaning of Gambling Estimate: Compression and Decompression Using Identical Twins

It has been asserted that the gambling approach elicits both an estimate of the true probability of the text sequence as well as an estimate of the entropy of the ensemble of English from which the sequence was drawn. In this section we investigate the operational significance of gambling in terms of data compression. Specifically, we shall argue that if the text in question results in capital S_n, then $\log_2 S_n$ bits can be saved in a naturally associated deterministic data compression scheme. We shall further

assert that if the gambling is optimal, then the data compression achieves the Shannon limit.

We shall make the assumption that there is an identical twin to the gambler who will be receiving some encoding of the text. This identical twin is assumed to have precisely the same thought processes as the encoder. (See also the Shannon twin [1].)

The scheme we shall describe is essentially the Elias coding scheme for stochastic processes with respect to the distribution $p(x(n)) \triangleq 2^{-n}S(x(n))$, where we have set $x(n) = (x_1, x_2, \cdots, x_n)$. (See Elias's unpublished manuscript and Jelinek's discussion of Elias's scheme [105].)

Consider the following data compression algorithm that maps the text $x_1, x_2, \cdots, x_n$ into a code sequence $c_1, c_2, \cdots, c_k$, where $c_i \in \{0,1\}$, $x_i \in \{0,1\}$, $i = 1,2,\cdots$. (We have assumed the text to be binary, without loss of generality, to obviate certain notational problems concerning bases of logarithms, etc.) Both the compressor and the decompressor know n. Let the 2^n text sequences be arranged in lexicographical order. Thus for example $0100101 < 0101101$. The encoder observes the sequence $x(n) = (x_1, x_2, \cdots, x_n)$. He then inspects his mental processes to calculate what his capital $S_n(x'(n))$ would have been on all sequences $x'(n) \leqslant x(n)$ and calculates $F(x(n)) = \sum_{x'(n) \leqslant x(n)} 2^{-n} S_n(x'(n))$. Clearly $F(x(n)) \in [0, 1]$. Let $k = \lceil n - \log S_n(x(n)) \rceil \triangleq \lceil -\log p(x(n)) \rceil$. Now express $F(x(n))$ as a binary decimal to k place accuracy: $\lfloor F(x(n)) \rfloor = c_1 c_2 \cdots c_k$. The sequence $c(k) = (c_1, c_2, \cdots, c_k)$ is transmitted to the decoder.

The decoder twin can calculate the precise $S(x'(n))$ associated with each of the 2^n sequences $x(n)$. He thus knows the cumulative sum of $2^{-n}S(x'(n))$ up through any sequence $x(n)$. He tediously calculates this sum until it first exceeds $.c(k)$. The first sequence $x(n)$ such that the cumulative sum falls in the interval $[.c_1 \cdots c_k, .c_1 \cdots c_k + (1/2)^k)$ is uniquely defined, and the size of $S(x(n))/2^n$ guarantees that this sequence will be precisely the encoded $x(n)$. Thus the twin has uniquely recovered $x(n)$. The number of bits required is $k = \lceil n - \log S(x(n)) \rceil$. The number of bits saved is $n - k = \lfloor \log S(x(n)) \rfloor$. For proportional gambling, $S(x(n)) = 2^n p(x(n))$; thus $Ek = \sum p(x(n)) \lceil -\log p(x(n)) \rceil \leqslant H(X_1, \cdots, X_n) + 1$. (An encoding-decoding algorithm for optimal data compression using these ideas and requiring only two operations per bit has been developed by Pasco [106].)

We see that if the betting operation is deterministic and is known both to the encoder and the decoder, then the number of bits necessary to encode $x_1, \cdots, x_n$ is approximately $n - \log S_n$ and that the expected value of this quantity is $H(X_1, \cdots, X_n)$. Thus for the text used in this experiment we argue that the gambling results correspond precisely to the data compression that would have been achieved by the given human encoder-decoder indentical twin pair.

In Section VII on the evaluation of experimental results, we see that the possibility of the identical twin encoding-decoding scheme applies in Section VII to 1) the average capital scheme where now we need an identical twin committee on the other end; 2) the best subject estimate scheme where now we need an extra $\log m$ bits of information to specify which of the gamblers is used for decoding; and 3) the committee gambling scheme where we have an identical committee on the other end. Thus the computed entropy estimates correspond to the actual compressions that are achieved.

TABLE I

EXPERIMENTAL RESULTS ON ESTIMATING ENTROPY OF ENGLISH USING SEQUENCE OF 75 SYMBOLS FROM *Jefferson the Virginian*

Subject	Capital Achieved	Resultant Entropy Estimate
1	1.50×10^{78}	1.29 bits/sym
2	1.46×10^{76}	1.38
3	3.36×10^{75}	1.41
4	2.37×10^{73}	1.51
5	6.45×10^{71}	1.57
6	3.22×10^{71}	1.59
7	2.30×10^{70}	1.64
8	4.00×10^{70}	1.67
9	2.21×10^{69}	1.68
10	9.63×10^{68}	1.70
11	3.88×10^{67}	1.76
12	3.60×10^{64}	1.90

Average capital achieved: 1.28×10^{77}.
Average capital estimate: 1.34.

VI. EXPERIMENTAL RESULTS

The above gambling procedure was carried out using twelve subjects and a sample of text from the same source Shannon used, *Jefferson the Virginian*, by Dumas Malone.[2] The sample of text used is given in the Appendix. Table I shows the resulting entropy estimates.

One disadvantage of this type of entropy estimate lies in the time necessary to perform the experiment. Each sample point in Table I was found by having each subject work interactively at a computer terminal for a period of approximately five hours. Each subject was allowed to augment his knowledge of English with digram and trigram statistics. We found, however, that the best estimates came from subjects who did not use the tables as a crutch. Subjects were also allowed to read as much of the book as they wished up to the sample in question in order to familiarize themselves with the author's style. Although each subject was tested separately, there was a definite air of competition.

Under the assumption that a more current piece of literature relating more directly to the subjects involved in the experiment might give a better estimate, *Contact: The First Four Minutes*, by Leonard and Natalie Zunin, was chosen as a second text source. The passage used appears in the Appendix. The results from two subjects are given in Table II.

[2]Awarded Pulitzer Prize in 1974 for his five volume series, *Jefferson and His Time*.

TABLE II
EXPERIMENTAL RESULTS ON ESTIMATING ENTROPY OF ENGLISH USING SEQUENCE OF 220 SYMBOLS FROM *Contact*

Subject	Capital Achieved	Resultant Entropy Estimate
1	5.62×10^{231}	1.26 bits/sym
2	6.01×10^{228}	1.30

VII. EVALUATION OF EXPERIMENTAL RESULTS

If only one experimental subject is available, the $\hat{H}=(1-(1/n)\log_{27} S_n)\log_2 27$ is the natural estimate of the entropy, as argued previously. Now we consider natural methods of combining the performance of m experimental subjects in order to obtain a better estimate of $H(X)$. Assume m remains fixed. There are two sources of error.

1) Bias—a subject may use an "incorrect" $p(x(n))$.
2) Statistical error—the sequence $x(n)$ may not be typical of the process; i.e., $-(1/n)\log_2 p(x(n))$ may differ significantly from $H(X)$.

The first source of error is handled by convexity, and the second by the asymptotic equipartition property.

Let subject i, $i=1,2,\cdots,m$, use gambling scheme $b_i(x(n))$, thus accumulating capital $S_n^{(i)}=b_i(x(n))27^n$. Consider the following four natural estimates for S_n and $H(X)$. In the first three of these, $\hat{H}=(1-(1/n)\log_{27} S_n)\log_2 27$.

a) *Average Capital*:

$$\bar{S}_n=\frac{1}{m}\sum_{i=1}^{m} S_n^{(i)}.$$

This is equivalent to a gambling scheme

$$b_{\text{avg}}(x(n))=\sum_{i=1}^{m}\frac{1}{m}b_i(x(n)); \tag{21}$$

i.e., each gambler begins with $(1/m)$th unit initial capital.

b) *Best Subject Estimate*:

$$S_n=\max_{i\in\{1,2,\cdots,m\}} S_n^{(i)}. \tag{22}$$

c) *Committee Gambling*:

$$b(x_k|x(k-1))=\sum_{i=1}^{m}\alpha_k^{(i)}b_i(x_k|x(k-1)) \tag{23}$$

where $\sum_{i=1}^{m}\alpha_n^{(i)}=1$, $\alpha_1\in[0,1]$, $i=1,\cdots,m$, and $S_n=27^n\prod_{k=1}^{n}b(x_k|x(k-1))$.

d) *Average Entropy Estimate*:

$$\hat{H}=\frac{1}{m}\sum_{i=1}^{m}\left(1-\frac{1}{n}\log_{27} S_n^{(i)}\right)\log_2 27=\frac{1}{m}\sum_{i=1}^{m}\hat{H}^{(i)}. \tag{24}$$

We reject d) immediately because it is too sensitive to poor gambling schemes on the part of one or more of the subjects. Suppose for example that b_1 bets all of his capital on one symbol at time 1 and loses. Then $S_n^{(1)}\equiv 0$, $n=1,2,\cdots$, and $(1/n)\log_{27} S_n^{(1)}=-\infty$, for all n, thus yielding $\hat{H}=+\infty$. This is an absurd use of the data.

Suppose that subject i achieves a limit $H^{(i)}$, i.e.,

$$\left[1-\frac{1}{n}\log_{27} S_n^{(i)}\right]\log_2 27\to H^{(i)}. \tag{25}$$

Thus $H^{(i)}$ is his asymptotic estimate of the entropy.

We now show that a) and b) both yield $H=\min_i H^{(i)}$ as the asymptotic estimate of H. Without loss of generality let

$$H^{(1)}<H^{(i)}, \qquad \text{for all } i. \tag{26}$$

Note that in a)

$$\begin{aligned}&\left(1-\frac{1}{n}\log_{27}\bar{S}_n\right)\log_2 27\\ &=\left(1-\frac{1}{n}\log_{27}\frac{1}{m}\sum_{i=1}^{m}S_n^{(i)}\right)\log_2 27\\ &=\left(1-\frac{1}{n}\log_{27}\left(S_n^{(1)}\right)-\frac{1}{n}\log_{27}\frac{1}{m}\sum_{i=1}^{m}\left(S_n^{(i)}/S_n^{(1)}\right)\right)\\ &\quad\cdot\log_2 27\to H^{(1)}\end{aligned} \tag{27}$$

since the last term $\to 0$, by (25) and (26). Thus a) and b) have the same limit for a fixed number of subjects m and for $n\to\infty$.

A similar argument can be formulated in the case of committee gambling for the special case of

$$\alpha_k^{(i)}=\alpha_k^{(i)}(x(k-1))=\frac{S_k^{(i)}(x(k-1))}{\sum_{i=1}^{m}S_k^{(j)}(x(k-1))}.$$

This corresponds to a weighted average of several betting schemes where the weighting factor is the proportion of money won by the ith scheme at time k. From c) we see

$$\begin{aligned}b(x_{k+1}|x(k))&=\sum_{i=1}^{m}\frac{S_k^{(i)}}{\sum_{j=1}^{m}S_k^{(i)}}b_i(x_{k+1}|x(k))\\ &=\sum_{i=1}^{m}\frac{b_i(x(k))}{\sum_{j=1}^{m}b_j(x(k))}b_i(x_{k+1}|x(k))\\ &=\frac{\frac{1}{m}\sum_{i=1}^{m}b_i(x(k+1))}{\frac{1}{m}\sum_{j=1}^{m}b_j(x(k))}\\ &=\frac{b_{\text{avg}}(x(k+1))}{b_{\text{avg}}(x(k))}=b_{\text{avg}}(x_{k+1}|x(k))\end{aligned} \tag{28}$$

where b_{avg} is the gambling scheme resulting in an average capital estimate in (21). Thus a), b), and c) are all equivalent in the special case when the weighting factor $\alpha_n^{(i)}$ is proportional to the current capital earned by the ith gambler.

In general, any other linear combination is possible and may do better than a) or b). As an example consider $\alpha_n^{(i)}=1/m$. Using the data in Table I and the conditional probability distributions guessed at each point by each of the 12 participants we arrive at $\hat{H}=1.25$, a lower estimate

TABLE III
EVALUATION OF ENTROPY ESTIMATES FOR DATA USED TO PRODUCE TABLE I

a) Average Capital Estimate:

$$\bar{S}_{75} = 1.28 \times 10^{77} \; ; \quad \hat{H} = 1.34$$

b) Best Subject Estimate:

$$S_{75} = 1.50 \times 10^{78} \; ; \quad \hat{H} = 1.29$$

c) Committee Gambling Estimate:

$$\alpha_n^{(i)} = 1/12 \; ; \; S_{75} = 1.24 \times 10^{79} \; ; \; \hat{H} = 1.25$$

d) Average Entropy Estimate (a rejected method):

$$\hat{H} = 1.59$$

than the best subject achieved. However, any choice of $\alpha_n^{(i)}$ yields an estimate the expectation of which is an upper bound to $H(X)$.

A summary of all of the above schemes as applied to the data used in calculating Table I is given in Table III.

VIII. CONCLUSIONS

Using the committee decision estimate as the estimate of the entropy of printed English, we discover a redundancy of at least 64 percent. The gambling winnings leading to this estimate have a direct data compression interpretation (Section V). Thus the ability of the experimental subjects to quantify their predictions would enable them to describe the given text in 36 percent of the original length.

The results of this paper also apply to the complexity of images, music, and computer programs.

ACKNOWLEDGMENT

The authors would like to thank the many students and faculty at Stanford who donated their time in helping the authors acquire experimental data for this paper. In particular we would like to thank Prof. John T. Gill, III, who achieved the highest capital growth rate in our experiments. Finally, we wish to thank the referees for mentioning Kolmogorov's contribution and the reference texts [64], [98].

APPENDIX

Excerpt from *Jefferson the Virginian*, by Dumas Malone (test section given in footnote).

> The surviving descriptions of her are meager, and there is none contemporary with these events. In comparison with him, she certainly was not tall; as an old slave put it, she was "low." The tradition is that her figure was slight, though well-formed, that she had large hazel eyes and luxuriant auburn hair. Within the family much was said afterwards about her beauty, and this can be accepted in essence though not in full detail.[18] Jefferson himself was straight and strong and his countenance was not unpleasing, but he was not a handsome man; beyond a doubt he prided himself on winning a pretty wife. There is considerable evidence of her amiability and her sprightliness of manner.[19] Her gaiety of spirit offset the characteristic seriousness of her lover; in her presence he could unbend. Gentle and sympathetic people always attracted him most, and clearly she was that sort, though she may have had her fiery moments before childbearing wore her out.
>
> She was not only a "pretty lady" but an accomplished one in the customary ways, and her love for music was a special bond with him. She played on the harpsichord and the pianoforte, as he did on the violin and the cello. The tradition is that music provided the accompaniment for his successful suit: his r9vals are said to have departed in admitted defeat after hearing him play and sing with her.[20] In later years he had the cheerful habit of singing and humming to himself as he went about his plantation. This is not proof in its[3]

Excerpt from *Contact*, by Leonard and Natalie Zunin (test section given in footnote.)

> A handshake refused is so powerful a response that most people have never experienced it or tried it. Many of us may have had the discomfort of a hand offered and ignored because it was not noticed, or another's hand was taken instead. In such an event, you quickly lower your hand or continue to raise it until you are scratching your head, making furtive glances to assure yourself that no one saw! When tw[4]

REFERENCES

[1] C. E. Shannon, "Prediction and entropy of printed English," *Bell Syst. Techn. J.*, pp. 50–64, Jan. 1951.
[2] V. Maixner, "Some remarks on entropy prediction of natural language texts," *Inform. Stor. Retr.*, vol. 7, pp. 293–295, 1971.
[3] A. P. Savchuk, "On the evaluation of the entropy of language using the method of Shannon," *Theory Prob. Appl.*, vol. 9, no. 1, pp. 154–157, 1964.
[4] T. Nemetz, "On the experimental determination of the entropy," *Kybern.* 10, pp. 137–139, 1972.
[5] G. P. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory Prob. Appl.*, vol. 4, no. 3, pp. 333–336, 1959.
[6] E. Pfaffelhuber, "Error estimation for the determination of entropy and information rate from relative frequencies," *Kybern.* 8, pp. 50–51, 1971.
[7] C. R. Blyth, "Note on estimating information," Tech. Rep. 17, Dept. of Statistics, Stanford Univ., Stanford, CA, 1958.
[8] G. A. Barnard, "Statistical calculation of word entropies for four western languages," *IRE Trans. Inform. Theory*, no. 1, pp. 49–53, 1955.
[9] Leonard J. Savage, "Elicitation of personal probabilities," *J. Amer. Statist. Ass.*, vol. 66, no. 336, pp. 783–801, Dec. 1971.
[10] David H. Bailey, "Sequential schemes for classifying and predicting ergodic processes," Ph.D. thesis, Stanford Univ., Stanford, CA, 1976.
[11] Noam Chomsky, "Three models for the description of language," *IRE Trans. Inform. Theory*, IT-2, no. 3, pp. 113–124, 1956.
[12] Benoit Mandelbrot, "An informational theory of the statistical structure of language," *Communication Theory*, W. Jackson, Ed. New York: Academic, 1953, pp. 485–502.
[13] J. Berry, "Some statistical aspects of conversational speech," *Communication Theory*, W. Jackson, Ed. New York: Academic, 1953, pp. 392–401.
[14] D. A. Bell, "The internal information of English words," *Communication Theory*, W. Jackson, Ed. New York: Academic, 1953, pp. 383–391.
[15] Victor H. Yngve, "Gap analysis and syntax," *IRE Trans. Inform. Theory*, IT-2, no. 3, pp. 106–112, 1956.
[16] E. G. Newman and L. J. Gerstman, "A new method for analyzing

[3]Test sequence: elf that he was a pleasing vocal performer but with Martha in the parlor it

[4]Test sequence: o people want to shake our hand simultaneously we may grab both one in a handshake and the other in a kind of reverse twist of the left hand which serves very well as a sign of cordiality and saves someone embarrassment.

printed English," *J. Exp. Psych.*, vol. 44, pp. 114–125, 1952.
[17] M. Grignetti, "A note on the entropy of words in printed English," *Inform. Contr.*, vol. 7, pp. 304–306, 1964.
[18] N. G. Burton and J. C. R. Licklider, "Long-range constraints in the statistical structure of printed English," *Amer. J. Psych.*, no. 68, pp. 650–653, 1955.
[19] W. J. Paisley, "The Effects of authorship, topic structure, and time of composition on letter redundancy in English texts," *J. Verbal Learn. Behav. 5*, pp. 28–34, 1966.
[20] A. Treisman, "Verbal responses and contextual constraints in language," *J. Verbal Learn. Behav.* 4, pp. 118–128, 1965.
[21] G. R. Miller and E. B. Coleman, "A set of thirty-six prose passages calibrated for complexity," *J. Verbal Learn. Behav.* 6, pp. 851–854, 1967.
[22] H. E. White, "Printed English compression by dictionary encoding," *Proc. IEEE*, vol. 55, no. 3, pp. 390–396, Mar. 1967.
[23] D. Jamison and K. Jamison, "A note on the entropy of partially-known languages," *Inform. Contr.*, vol. 12, pp. 164–167, 1968.
[24] K. R. Rajagopalan, "A note on entropy of Kannada prose," *Inform. Contr.*, vol. 8, pp. 640–644, 1965.
[25] E. Newman and N. Waugh, "The redundancy of texts in three languages," *Inform. Contr.*, vol. 3, pp. 141–153, 1960.
[26] G. Siromoney, "Entropy of Tamil prose," *Inform. Contr.*, vol. 6, pp. 297–300, 1963.
[27] P. Balasubrahmanyam and G. Siromoney, "A note on entropy of Telugu prose," *Inform. Contr.*, vol. 13, pp. 281–285, 1968.
[28] M. A. Wanas, A. I. Zayed, M. M. Shaker, and E. H. Taha, "First-second- and third-order entropies of Arabic text," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, p. 123, Jan. 1976.
[29] H. Hansson, "The entropy of the Swedish language," *Trans. Second Prague Conf. on Inform. Theory, Statistical Decision Functions, and Random Processes*, Prague, 1960, pp. 215–217.
[30] B. S. Ramakrishna, K. K. Nair, V. N. Chiplunkar, B. S. Atal, V. Ramachandran, and R. Subramanian, "Relative efficiencies of Indian languages," *Nature*, vol. 189, no. 4768, pp. 614–617, 1961.
[31] B. S. Ramakrishna and R. Subramanian, "Relative efficiency of English and German languages for communication of semantic content," *IRE Trans. Inform. Theory*, vol. 4, no. 3, pp. 127–129, 1958.
[32] N. Tzannes, V. Spencer, and A. Kaplan, "On estimating the entropy of random fields," *Inform. Contr.*, vol. 16, pp. 1–6, 1970.
[33] J. R. Parks, "Prediction and entropy of half-tone pictures," *Behavioral Science*, vol. 10, pp. 436–445, 1965.
[34] G. A. Miller and F. C. Frick, "Statistical behavioristics and sequences of responses," *Psych. Rev.*, vol. 56, pp. 311–324, 1949.
[35] G. A. Miller and J. A. Selfridge, "Verbal context and the recall of meaningful material," *Amer. J. Psych.*, vol. 63, pp. 176–185, 1950.
[36] E. G. Newman, "Computational methods useful in analyzing series of binary data," *Amer. J. Psych.*, vol. 64, pp. 252–262, 1951.
[37] E. B. Newman, "The pattern of vowels and consonants in various languages," *Amer. J. Psych.*, vol. 64, pp. 369–379, 1951.
[38] F. C. Frick and G. A. Miller, "A statistical description of operant conditioning," *Amer. J. Psych.*, vol. 64, pp. 20–36, 1951.
[39] A. Chapanis, "The reconstruction of abbreviated printed messages," *J. Experimental Psych.*, vol. 48, no. 6, pp. 496–510, 1954.
[40] W. F. Bennett, P. M. Fitts, and M. Noble, "The learning of sequential dependencies," *J. Experimental Psych.*, vol. 48, no. 4, pp. 303–312, 1954.
[41] G. A. Miller, E. B. Newman, and E. A. Friedman, "Length-frequency statistics for written English," *Inform. Contr.*, vol. 1, pp. 370–389, 1958.
[42] D. H. Carson, "Letter constraints within words in printed English," *Kybern.*, vol. 1, pp. 46–54, 1961.
[43] C. P. Bourne and D. F. Ford, "A study of the statistics of letters in English words," *Inform. Contr.*, vol. 4, pp. 48–67, 1961.
[44] J. A. Hogan, "Copying redundant messages," *J. Experimental Psych.*, vol. 62, no. 2, pp. 153–157, 1961.
[45] N. M. Blachman, "Prevarication versus redundancy," *Proc. IRE*, pp. 1711–1712, 1962.
[46] E. Tulving, "Familiarity of letter-sequences and tachistoscopic identification," *Amer. J. Psych.*, vol. 76, pp. 143–146, 1963.
[47] R. N. Shepard, "Production of constrained associates and the informational uncertainty of the constraint," *Amer. J. Psych.*, vol. 76, pp. 218–228, 1963.
[48] E. S. Schwartz, "A dictionary of minimum redundancy encoding," *J. Ass. Comput. Mach.*, vol. 10, pp. 413–439, 1963.
[49] H. Bluhme, "Three-dimensional crossword puzzles in Hebrew," *Inform. Contr.*, vol. 6, pp. 306–309, 1963.
[50] P. H. Tannenbaum, F. Williams, and C. S. Hillier, "Word predictability in the environments of hesitations," *J. Verbal Learn. Behav.*, vol. 4, pp. 134–140, 1965.
[51] J. Raviv, "Decision making in Markov chains applied to the problem of pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-3, no. 4, Oct. 1967.
[52] E. S. Schwartz, and A. J. Kleiboemer, "A language element for compression coding," *Inform. Contr.*, vol. 10, pp. 315–333, 1967.
[53] R. B. Thomas, and M. Kassler, "Character recognition in context," *Inform. Contr.*, vol. 10, pp. 43–64, 1967.
[54] R. W. Cornew, "A statistical method of spelling correction," *Inform. Contr.*, vol. 12, pp. 79–93, 1968.
[55] D. McNicol, "The confusion of order in short-term memory," *Austr. J. Psych.*, vol. 23, no. 1, pp. 77–84, 1971.
[56] J. J. Tuinman and G. Gray, "The effect of reducing the redundancy of written messages by deletion of function words," *J. Psych.*, vol. 82, pp. 299–306, 1972.
[57] B. Mandelbrot, "Simple games of strategy occurring in communication through natural languages," *IRE Trans. Inform. Theory*, vol. PGIT-3, pp. 124–137, 1954.
[58] W. F. Schreiber, "The measurements of third order probability distributions of television signals," *IRE Trans. Inform. Theory*, vol. IT-2, no. 3, pp. 94–105, 1956.
[59] J. O. Limb, "Entropy of quantised television signals," *Proc. IEEE* vol. 115, no. 1, pp. 16–20, 1968.
[60] R. C. Pinkerton, "Information theory and melody," *Sci. Amer.*, vol. 194, no. 2, pp. 77–86, 1956.
[61] J. E. Youngblood, "Style as information," *J. Music Theory*, vol. 2, no. 1, p. 24, 1958.
[62] J. E. Cohen, "Information theory and music," *Behav. Sci.*, vol. 7, no. 2, pp. 137–163, 1962.
[63] K. R. Siromoney and K. R. Rajagopalan, "Style as information in Karnatic music," *J. Music Theory*, vol. 8, no. 2, pp. 267–272, 1964.
[64] A. M. Yaglom and J. M. Yaglom, *Probability and Information*, 3rd revised ed. Leningrad: Science House, 1973, Chapter IV, Part 3, pp. 236–329.
[65] R. G. Piotrovski, *Informational Measurements of Language*, Leningrad: Nauka (Science), 1961, pp. 79–87.
[66] A. M. Yaglom, J. M. Yaglom, and R. L. Dobrushin, "The theory of information and linguistics," *Questions of Linguistics*, pp. 100–110, 1960.
[67] R. L. Dobrushin, "Mathematical methods in linguistics," *Math. Ed.* (new series), no. 6, Moscow: Phizmattiz, 1961, pp. 37–60.
[68] V. Belevich, "The theory of information and linguistic statistics," *Bulletin of Royal Academy of Belgium*, (Category of Sciences), pp. 419–436, 1956.
[69] V. Yu. Urbach, "Taking into account correlations between letters of the alphabet when computing the information content of a message," *Prob. Cybern.*, no. 10, pp. 111–117, 1963.
[70] G. Siromoney, "An information-theoretical test for familiarity with a foreign language," *J. Experimental Psych.*, no. 8, pp. 1–6, 1964.
[71] P. B. Nevel'skii and M. D. Rosenbaum, "Guessing the professional texts of specialists and nonspecialists," *Statistics of Speech and Automatic Textual Analysis*, Leningrad: Nauka (Science), 1971, pp. 134–148.
[72] K. Küpfmüller, "The entropy of the German language," *Fernmeldtechnische Zeitschrift* (J. of Commun.), no. 6, pp. 265–272, 1954.
[73] N. Petrova, R. Piotrovski, and R. Giraud, "The entropy of written French," *Bulletin of Society of Linguistics of Paris*, vol. 58, no. 1, pp. 130–152, 1964.
[74] R. Manfriono, "The entropy of the Italian language and its computation," *Alta Frequenza* (High Frequency), vol. 29, no. 1, pp. 4–29, 1960.
[75] L. Dolezel, "Predictions of the Entropy and redundancy of Czechoslovakian texts," *Slovo a Sbovesnost*, vol. 24, no. 3, pp. 165–175, 1963.
[76] F. Zitek, "Some comments on the entropy of the Czechoslovakian language," *Trans. Prague Conf. Inform. Theory, Statistical Decision Functions, and Random Processes*, Prague, 1964, pp. 841–846.
[77] E. Nicolau, C. Sala, and A. Roceric, "Observations on the entropy of the Rumanian language," *Studisi Cercetai Linvist*, vol. 10, no. 1, pp. 35–54, 1959.

[78] R. A. Kazarian, "The evaluation of the entropy of an Armenian text," *News of the Academy of Sciences of Armenia* (*The Physical and Mathematical Sciences*), vol. 14, no. 41, pp. 161–173, 1961.
[79] D. N. Lenskii, "On the evaluation of the entropy of Adegei printed texts," *Scientific Notes of the Kabardino-Balkarskii Univ.* (*The Physical and Mathematical Series*), issue 16, Nal'chik, pp. 165–166, 1962.
[80] T. I. Ibragimov, "An evaluation of the mutual information of letters in the Tartar language," *Scientific Notes of the Univ. of Kazan*, vol. 124, book 2, Kazan, pp. 141–145, 1964.
[81] N. Rychkova, "Linguistics and mathematics," *Nauka i Zhizn* (Science and Life), no. 9, pp. 76–77, 1961.
[82] P. M. Alekseev, "Frequency count dictionaries of English and their practical application," *Statistics of Speech and Automatic Textual Analysis*, Leningrad: Nauka (Science), 1971, pp. 160–178.
[83] G. A. Miller, "Speech and language," *Experimental Psych.*, S. S. Stevens, Ed., vol. 2, published in translation by IL, Moscow, 1963, pp. 348–374.
[84] B. Mandelbrot, "An informational theory of the statistical structure of language," in *Communications Theory*, by W. Jackson. New York: Academic, 1953, pp. 486–502.
[85] A. A. Piotrovskaia, R. G. Piotrovskii, and K. A. Razzhivin, "The entropy of the Russian language," *Questions of Linguistics*, no. 6, pp. 115–130, 1962.
[86] O. L. Smirnov and A. V. Ekimov, "The entropy of Russian telegraph texts," *Projects of the Leningrad Institute of Aviation Instrument Construction*, no. 54, (systems for processing and transmitting information), Leningrad, 1967, pp. 76–84.
[87] T. Tarnoczy, "On factors creating differences in the entropy of a language," *Nyelvtudomanyi Kozlemenyek*, no. 63, pp. 161–178, 1961.
[88] A. M. Kondratov, "The theory of information and poetics (The entropy of the rhythm of the Russian language)," *Probl. Cybern.*, no. 9, pp. 279–286, 1963.
[89] S. Marcus, "Entropy and Poetical energy," *Notes on Theoretical and Applied Linguistics*, no. 4, pp. 171–180, 1967.
[90] H. Kreuzer and R. Gunzenhauser, *Mathematics and Poetry*, Nympfenburger Verlaghandlung, Munich, 1968.
[91] W. Endres, "A comparison of the redundancy in written and spoken language," *Proc. 2nd International Conf. for Inform. Theory*, Tsakhadzor, Armenia, 1971.
[92] A. Fradis, L. Mihailescu, and I. Voinescu, "The entropy and informational energy of the spoken Rumanian language," *Rumanian J. Linguistics*, vol. 12, no. 4, pp. 331–339, 1967.
[93] I. Voinescu, A. Fradis, and L. Mihailescu, "The first degree entropy of phonemes in aphasics," *Rumanian Review of Neurology*, vol. 4, no. 1, pp. 67–79, 1967.
[94] M. Roland, "An investigation using musical examples of the decrease in entropy due to interdependence between several sources of information, and extensions to higher order markov chains," *Research Rep. of the State of Nordheim-Westfalen*, no. 1768, pp. 39, 41, 43–44, 79–80, 1967.
[95] D. S. Lebedev and I. I. Tsukkerman, *Television and the theory of information*, Energiia (Energy), Moscow, 1965.
[96] D. S. Lebedev and Piil', "Experimental Research in the statistics of television messages," *Tekhnika Kino i Televideniia* (Technology of Films and Television), no. 3, pp. 37–39, 1959.
[97] G. A. Kaiser, "On the entropy of typewritten text," *Nachrichtentechn. Zeitschrift*, vol. 13, no. 5, pp. 219–224, 1960.
[98] K. Weltner, *The measurement of verbal information in psychology and education*, Springer-Verlag, 1973.
[99] A. M. Zubkov, "Limit distribution for a statistical estimate of entropy," *Teorija Verojatnost i Primenen*, no. 18, pp. 643–650, 1973.
[100] G. A. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology*, H. Quastler, Ed. New York: Free Press, 1955.
[101] B. Harris, "The statistical estimation of entropy in the nonparametric case," *Topics in Information Theory*, I. Csiszar and P. Elias, Ed. (Colloquia Mathematica Societatis Janos Brelyai, 16), North-Holland, 1977, pp. 323–357.
[102] J. Kelly, Jr., "A new interpretation of information rate," *Bell Syst. Tech. J.*, pp. 917–926, July 1956.
[103] L. Breiman, "Optimal gambling systems for favorable games," *Proc. Fourth Berkeley Sym.*, vol. 1, pp. 65–78, 1961.
[104] T. Cover, "Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin," Tech. Rep. no. 12, Statistics Dept., Stanford Univ., Stanford, CA, Oct. 1974, to appear *Ann. Stat.*
[105] F. Jelinek, *Probabilistic Information Theory*, New York: McGraw-Hill, 1968.
[106] R. C. Pasco, "Source coding algorithms for fast data compression," Ph.D. thesis, Dep. Electrical Engineering, Stanford Univ., Stanford, CA, May 1976.