Pergamon

*Chaos, Solitons & Fractals* Vol. 4, No. 5, pp. 709–743, 1994

0960-0779/94$7.00 + .00

0960–0779(93)E0002-7

# Entropy of Natural Languages: Theory and Experiment

LEV B. LEVITIN

College of Engineering, Boston University, 44 Cummington St, Boston, MA 02215, USA

**and**

ZEEV REINGOLD

Bezeq Telecommunications Corp., Tel-Aviv, Israel

(*Received* 15 *December* 1992; *revised and accepted for publication* 25 *August* 1993)

**Abstract**—The concept of the entropy of natural languages, first introduced by Shannon [A mathematical theory of communications, *Bell Syst. Tech. J.* **27**, 379–423 (1948)] and its significance is discussed. A review of various known approaches to and results of previous studies of language entropy is presented. A new improved method for evaluation of both lower and upper bounds of the entropy of printed texts is developed. This method is a refinement of Shannon's prediction (guessing) method [Shannon, Prediction and entropy of printed English, *Bell Syst. Tech. J.* **30**, 50–64 (1951)]. The evaluation of the lower bound is shown to be a classical linear programming problem. Statistical analysis of the estimation of the bounds is given and procedures for the statistical treatment of the experimental data (including verification of statistical validity and significance) are elaborated. The method has been applied to printed Hebrew texts in a large experiment (1000 independent samples) in order to evaluate entropy and other information-theoretical characteristics of the Hebrew language. The results have demonstrated the efficiency of the new method: the gap between the upper and lower bounds of entropy has been reduced by a factor of 2.25 compared to the original Shannon approach. Comparison with other languages is given. Possible applications of the method are briefly discussed.

## NOTATION

| | |
|---|---|
| $L$ | length of a sequence of consecutive symbols ($L$-gram) |
| $K$ | number of different symbols in an alphabet of a discrete source (the size of the alphabet) |
| $s_j^L$ | particular $L$-gram ($j = 1, 2, \ldots, K^L$) |
| $\Omega = \{s_j^L\}$ | set of all the possible $L$-grams |
| $S^L$ | random variable associated with all the possible $L$-grams |
| $p_{S^L}(s_j^L) = \Pr\{S^L = s_j^L\}$ | probability of a particular sequence $s_j^L$ |
| $x_k$ | symbol of the alphabet ($k = 1, 2, \ldots, K$) |
| $k$ | alphabetic number of the symbol $x_k$ |
| $X$ | random variable associated with all the symbols of the alphabet |
| $P_{S^L,X}(s_j^L, x_k) = \Pr\{S^L = s_j^L, X = x_k\}$ | joint probability of an $L$-gram $s_j^L$ and the following symbol $x_k$ |
| $P_{X/S^L}(x_k/s_j^L) = \Pr\{X = x_k/S^L = s_j^L\}$ | conditional probability of a symbol $x_k$ following a given $L$-gram $s_j^L$ |
| $P_X(x_k) = \Pr\{X = x_k\}$ | marginal probability of a symbol $x_k$ |
| $F_{L+1} = F(X/S^L)$ | conditional entropy of $(L + 1)$th order of a text |
| $G_L(S^L)$ | joint $L$-gram entropy per symbol |
| $H(X)$ | entropy of a text per symbol |

| | |
|---|---|
| $F_0$ | zero-order entropy |
| $R$ | redundancy of a text |
| $I(S, X)$ | amount of information in the previous (infinite) text about the following symbol |
| $y_i$ $(i = 1, 2, \ldots, K)$ | symbol of the reduced text (the number of attempts) |
| $Y_L$ | random variable associated with symbols of the reduced text in an experiment of the prediction of a symbol, following an $L$-gram |
| $q_i = P_{Y_L}(y_i) = \Pr\{Y_L = y_i\}$ | probability of the number of attempts $y_i$ |
| $N$ | total number of trials in a guessing experiment |
| $m_i$ | number of trials in which the number of attempts was equal to $y_i$ |
| $\hat{q}_i = \dfrac{m_i}{N}$ | frequency of $y_i$ |
| $(E, \mathcal{B}, P)$ | probability space |
| $E = \{e_{j,k}\}$ | space of elementary events |
| $e_{j,k}$ | elementary event which is a sequence $s_j^L$ and a following letter $x_k$ |
| $\mathcal{B}$ | Borel field on $E$ |
| $P$ | probability measure on $E$ |
| $\mathbf{x}_k$ | element of Borel field $\mathcal{B}$ associated with the event that an $L$-gram is followed by symbol $x_k$ |
| $\mathbf{s}_j^L$ | element of the Borel field $\mathcal{B}$ associated with an $L$-gram $s_j^L$ |
| $\mathbf{y}_i$ | element of the Borel field $\mathcal{B}$ associated with a symbol of the reduced text $y_i$ |
| $k_i(j)$ | alphabetic number of a symbol occupying the $i$th place in the permutation of the alphabet determined by an $L$-gram |
| $p_i = P_{X/S^L}(x_{k_i(j)}/s_j^L)$ | conditional probability of the symbol $x_{k_j(j)}$ following a given $L$-gram $s_j^L$ (the ordered conditional probability) |
| $\underline{H}_{\mathrm{Sh}}(X)$ | Shannon's lower bound of entropy per symbol |
| $\bar{H}_{\mathrm{Sh}}(X) \underline{\Delta} H(Y_L)$ | Shannon's upper bound of entropy per symbol |
| $\hat{\bar{H}}_{\mathrm{Sh}}(X) = \hat{H}(Y_L)$ | estimate of Shannon's upper bound |
| $\hat{\underline{H}}_{\mathrm{Sh}}(X)$ | estimate of Shannon's lower bound |
| $\hat{\underline{R}}$, $\hat{\bar{R}}$ | estimates of the lower and upper bounds of redundancy, respectively |
| $a_r$ $(r = 1, 2, \ldots)$ | disjoint subsets of $\Omega$ obtained by subdivision for the upper bound |
| $\omega_r = \Pr\{s_j^L \in a_r\}$ | probability of the subset $a_r$ |
| $q_i^{(r)}$ | probability of the number of attempts $y_i$ if $s_j^L \in a_r$ |
| $\bar{H}_r$ | Shannon's upper bound of entropy for the subset $a_r$ |
| $\hat{H}(X)$ | total upper bound of entropy obtained by use of subdivision |
| $\mathbf{p} = (p_1, \ldots, p_K)$ | vector of ordered conditional probabilities |
| $\mathbb{C}$ | simplex of all the ordered probability vectors $\mathbf{p}$ |
| $\mathbf{v}_g$ $(g = 1, \ldots, K)$ | vertices of simplex $\mathbb{C}$ |
| $\alpha_g$ | weight coefficients of a probability vector in the simplex $\mathbb{C}$ |
| $\mathbf{p_j}$ | vector of ordered conditional probabilities of the symbol following a given $L$-gram $s_j^L$ |
| $H(\mathbf{p_j})$ | entropy of the probability distribution given by the vector $\mathbf{p_j} = (p_j^1, \ldots, p_j^K)$ |
| $\mathbf{q} = (q_1, \ldots, q_K)$ | mean probability vector of the simplex $\mathbb{C}$ |
| $C$ | convex set in a $(K-1)$-dimensional space |
| $G$ | number of extreme points of the convex set $C$ |
| $\mathbf{u}_g$ $(g = 1, \ldots, G;\ G \geqslant K)$ | extreme points of the convex set $C$ |
| $C_r$ $(r = 1, 2, \ldots)$ | convex subset of simplex $\mathbb{C}$ |
| $G_r$ | number of extreme points of $C_r$ |
| $\mathbf{u}_{g_r}^{(r)}$ $(g_r = 1, \ldots, G_r;\ G_r \geqslant K)$ | extreme points of $C_r$ |
| $d_r$ $(r = 1, 2, \ldots)$ | disjoint subsets of $\Omega$, such that if $s_j^L \in d_r$, then $\mathbf{p_j} \in C_r$ |
| $\rho_r = \Pr\{s_j^L \in d_r\}$ | the probability of a subset $d_r$ |
| $\mathbf{q}^{(r)}$ | mean probability vector corresponding to the subset $d_r$ |
| $\alpha_{g_r}^{(r)}$ | weight coefficients of $\mathbf{q}^{(r)}$ expressed in terms of $u_{g_r}^{(r)}$ |
| $\underline{H}(X) = \sum_r \rho_r \underline{H}_r$ | lower bound of entropy per symbol obtained by use of subdivision; $\underline{H}_r$ is the lower bound for $d_r$ |
| $\hat{\bar{H}}_r$ | estimate of the upper bound $\bar{H}_r$ for a subset $a_r$ |

| | |
|---|---|
| $\hat{\bar{H}}(X)$ | estimate of the (total) upper bound obtained by subdivision |
| $N_r$ | number of trials, where $L$-grams belong to $a_r$ |
| $m_i^{(r)}$ | number of trials belonging to $a_r$ with the outcome $y_i$ |
| $\hat{\omega}_r$ | estimate of $\omega_r$ |
| $\hat{q}_i^{(r)}$ | estimate of $q_i^{(r)}$ |
| $V(Z)$ | variance of a random variable $Z$ |
| $E(Z)$ | expected value of a random variable $Z$ |
| $B$ | bias of an estimate |
| $\hat{\mathbf{q}}^{(r)}$ | estimate of $\mathbf{q}^{(r)}$ |
| $\hat{\rho}_r$ | estimate of $\rho_r$ |
| $\hat{H}(X) = \sum_r \hat{\rho}_r \hat{H}_r$ | estimate of the lower bound $\underline{H}(X)$, $\hat{H}$-estimate of the lower bound for subset $d_r$ |
| $\check{\mathbf{q}}^{(r)}$ | smoothed estimate of $\mathbf{q}^{(r)}$ |
| $\delta_{ri,r'i'}$ | Kronecker's $\delta$-function |
| $p_0 = 0.85$ | a threshold value of probability |
| $F$ | objective functional in a linear programming problem |
| $A \Delta B$ | $A$ equals $B$ by definition |
| $A \simeq B$ | $A$ is approximately equal to $B$ |
| $A \gtrsim B$ | $A$ is approximately equal to $B$ or greater than $B$ |

## 1. INTRODUCTION: THE CONCEPT OF THE ENTROPY OF A LANGUAGE AND ITS SIGNIFICANCE

A natural language gives us a remarkable example of a system used for the generating of long sequences of symbols—texts—which possess some exceptional properties. Being, on one hand, "natural", a text is specially prepared in each case in order to serve as a message containing some specific information. To meet this purpose a text evolving in time or in space should be "random", i.e. not completely predictable: a completely predictable (deterministic) process would not provide any new information. Thus, a text in a natural language is a realization of a random process (Shannon [1, 2]; Mandelbrot [3]; Yaglom *et al.* [4]; Herdan [5]). But, in contrast with other random processes which exist in nature, this random process was developed, modified and selected during a long period of evolution and "natural selection" being specially intended for meaningful communication between human beings. So, the texts in a natural language represent a random process of a very unusual sort—a naturally produced information-carrying process. The stochastic nature of this random process is of extreme complexity. For instance, the question of ergodicity of the process is still controversial. In our opinion, it is, to some extent, a matter of interpretation, since it cannot be checked experimentally by use of rather short "segments" of realizations of this random process, which are the only empirical data available to us. In any case, the statistical interdependence between elements of a text is very complicated and spread over long ranges. What has been said can explain the importance of information-theoretical study of natural languages. An information-theoretical approach is highly relevant here, since we have to deal with a process specially built and intended for information transmission. Therefore one may expect that the results of such a study can be meaningful for both linguistics and communication theory. On the other hand, though complicated, this information-carrying random process has a clear, well-distinguishable structure and appeals much to our experience and intuition. We may hope that a deeper investigation of this process can provide us with a better insight into the nature of other random processes which are typical for the functioning of living bodies and intellectual beings, par excellence. As A. N. Kolmogorov pointed out, the totality of texts is the richest and unique material for such a study, which can shed new light on general rules and laws of mental information processes.

The basic information-theoretical characteristic of a language is the entropy per symbol of the text. The formal definition of the entropy of a discrete random process will be given in Section 2. Entropy per symbol is one of the most important concepts having a profound triple meaning. Entropy is a measure of the variety of all the possible texts. If entropy per symbol is $H$ bits, then there exist approximately $2^{NH}$ various texts of the large length (number of symbols) $N$. Entropy is also a measure of the uncertainty associated (on average) with each new symbol of a consecutively produced text: if a long preceding text is given, then there exist approximately $2^{NH}$ possible continuations of length $N$, all of them being almost equiprobable. At the same time, entropy is a measure of information which is obtained (on average) when the uncertainty is eliminated by letting know the actual following symbol of the text. It means that a text of length $N$ contains $NH$ bits of information and requires $NH$ binary digits, if the totality of all the possible texts is encoded by a binary code in the most economical way (with the minimum of binary digits per symbol).

Some other important characteristics of a language can be introduced on the basis of the entropy concept. The relative reduction of the text length attainable by optimal coding expresses the redundancy of the language (see Section 2), while the difference between the first-order entropy (entropy of the probability distribution of symbols considered as independent) and the entropy of "infinite order" characterizes the interdependence between the symbols of the text, its predictability and noise immunity. Formal restrictions, imposed on special types of texts, such as verses, can be also expressed in terms of entropy and such a characterization was shown to be of typological importance in philology (Rychkova [6]; Kondratov [7]). The distribution of information along the text is non-uniform, and the entropy of specific types of linguistic situations can be a useful tool in the analysis of language as a communication system.

All the above-stated explains the interest displayed by many researchers during 40 years for the entropy of language. Section 2 of this paper presents a review of different known approaches to, and results of, the study of language entropy. Sections 3–8 are devoted to the development of a new improved method for evaluation of the entropy of a language and to application of this method to a new object—to printed Hebrew. Proofs of theorems and other parts of research are omitted here due to lack of room. For more comprehensive information see Reingold [8].

## 2. APPROACHES TO THE STUDY OF LANGUAGE ENTROPY

### 2.1. *Written language as a random process; entropy of a language*

The totality of texts written in a certain language can be considered as a set of realizations of a random process generated by a certain source, where a "source" is a collective name for all the producers of the texts. As was mentioned before, this random process is of extremely complex stochastic nature. Being treated on the level of separate symbols it is a discrete random process produced by a source symbol-by-symbol, with an alphabet including letters and space (punctuation marks are usually ignored). This random process is assumed to be ergodic. Though it is by no means a finite-order Markovian process, a description of the language as a Markovian process of a high order can be used as a useful approximation.

Consider a sequence of consecutive symbols of length $L$ taken from a long text written in a given language. Such a sequence will be called an "$L$-gram". Denote the number of symbols in the alphabet by $K$; then there are $K^L$ possible $L$-grams, which we denote by $s_j^L$ ($j = 1, \ldots, K^L$).

Since we assume the text to be a realization of an ergodic process a random variable $S^L$ can be introduced which takes on values from the set $\Omega = \{s_j^L\}$ of all the possible $L$-grams with probabilities $P_{S^L}(s_j^L)$. The symbol $X$ following the random sequence $S^L$ is also a random variable, which takes on values $x_k$ from the alphabet $\{x_k\}$ $(k = 1, \ldots, K)$.

Denote by $P_{S^L,X}(s_j^L, x_k)$ the joint probability of the event that a symbol $x_k$ follows a $L$-gram $s_j^L$; by $P_{X/S^L}(x_k/s_j^L)$ the conditional probability of $x_k$ to follow $s_j^L$ and by $P_X(x_k)$ the marginal probability of the symbol $x_k$.

The conditional entropy of the random symbol $X$ following the random sequence $S^L$ is defined as

$$F_{L+1}(X/S^L) = -\sum_{s_j^L}\sum_{x_k} P_{S^L,X}(s_j^L, x_k) \ln P_{X/S^L}(x_k/s_j^L) \tag{2.1.1}$$

and measured in natural units-"nats". It is sometimes called "the entropy of the $(L + 1)$th order per symbol" of the random process. The conditional entropy $F_L$ is connected with the joint $L$-gram entropy per symbol

$$G_L = -\frac{1}{L}\sum_{s_j^L} P_{S^L}(s_j^L) \ln P_{S^L}(s_j^L) \tag{2.1.2}$$

by the following equation (Shannon [1], Theorem 6):

$$F_L = LG_L - (L - 1)G_{L-1}. \tag{2.1.3}$$

It was shown by Shannon [1] that $F_L$ is a monotonically decreasing function of $L$ and there exists a limit,

$$\lim_{L\to\infty} F_{L+1}(X/S^L) = H(X) \tag{2.1.4}$$

which is, by definition, the entropy of the random process (the printed language) per symbol. The $(L + 1)$th order entropy corresponds to the approximation of the random process by a Markov process of $L$th order.

According to the information-theoretical meaning of entropy, the entropy per symbol $H(X)$ characterizes the amount of information delivered (on average) by each symbol of the text. This amount of information is much less than the maximum possible amount of information per symbol $F_0$ which can be achieved by use of an alphabet of $K$ symbols provided the symbols are equiprobable and independent:

$$F_0 = \ln K. \tag{2.1.5}$$

Actually, the symbols in a natural language are far from being equiprobable, and therefore the first-order entropy $F_1$ defined on the base of unconditional probabilities of separate symbols (the dependence between consecutive symbols being neglected) is considerably less than the maximum entropy:

$$F_1(X) = -\sum_{x_k} P_X(x_k) \ln P_X(x_k) < F_0. \tag{2.1.6}$$

However, the main contribution to the difference between the maximum possible value of the entropy per symbol and the actual one is made by the stochastic dependence between the symbols. The statistical structure of a natural language is characterized by a "long-distance memory" so that even rather distant elements of the text can impose some additional constraints on the conditional probabilities of the current symbol. The effect of the memory is measured quantitatively by the amount of information in the preceding text

about the following symbol

$$I(S, X) \equiv \lim_{L\to\infty} I(S^L, X) = F_1(X) - H(X). \tag{2.1.7}$$

This information expresses the property of the noise immunity of the text: the more information the better a missing symbol can be restored on the ground of our knowledge of the previous text. The total influence of both factors—the non-equiprobability of the symbols and the dependence between them—can be characterized quantitatively by the redundancy $R$

$$R = \frac{F_0 - H(X)}{F_0}. \tag{2.1.8}$$

The value of redundancy is equal to the maximum possible compression of a text (reduction of the text length) provided we make a complete use of the statistical properties of the language. In other words, for a long text of $N$ symbols given in a natural language, the greatest lower bound $N_1$ of the length of an encoded text, obtained by use of a reversible coding, with an output alphabet of the same size, satisfies the relation

$$R = \frac{N - N_1}{N}. \tag{2.1.9}$$

This result is valid for any ergodic source and follows from the Shannon–McMillan–Breiman theorem (Shannon [1], Theorem 3).

The concept of redundancy finds wide applications in communication, data storage and cryptography.

A few years ago a paper was published (Hillberg [9]) which suggested a drastic revision of the entropy concept as applied to texts in natural languages. The author developed an abstract model of text structure that led him to a very strange conclusion: the total information does not grow linearly with the length of the text, but much slower, say, as a square root of the length. It means that the information rate tends very rapidly to zero with the length of the text. Taking into account the total length of texts already published, we would have to conclude that a new book of $10^6$ letters (about 500 pages) contains at most one bit of information only (?!) Nobody would be able to say anything new! To be exact, the author restricts (rather arbitrarily) the applicability of his results to "connected" or "related" ("zusammenhaengende") texts and sets "practical information values in the order of magnitude of $10^{-3}$–$10^{-4}$ bits per letter". Hillberg's own paper consists of approximately 20,000 letters. If his paper were considered from the standpoint of his own theory, it would contain no more than two information bits, which, in our opinion, makes his model inapplicable.

Several researchers (e.g. Ebeling and Nicolis [10]; Nicolis and Katsikas [11]) approached the problem of language entropy from the perspective of the nonlinear system dynamics. They developed interesting theoretical models which certainly add a new dimension to our understanding of the problem. However, in the words of Ebeling and Nicolis, "actually the present stage of the analysis does not even allow us to decide" which model describes correctly the real language phenomena.

## 2.2. *Approaches and results based on the statistics of languages*

A straightforward approach to the estimation of the entropy per symbol of a language is to use the statistics of $L$-grams in order to evaluate the $L$th order entropy. Such calculations were done first by Shannon [2] for English and then by many authors for a

number of various languages. In Table I, $L$-gram entropies for several languages are given.

It should be taken into account that the results obtained by Kuepfmueller [12] for German were based on the statistics of syllables and words rather than of $L$-grams, and the alphabet did not include the space. The results obtained for Hebrew by Ladany [13] were based on an alphabet of 28 symbols including the space and the five special end-forms of letters ("kaf", "mem", "nun", "peh", "tsadeh") being considered as separate letters. (The values of $F_L$ for Hebrew were calculated from the values of $G_L$ given originally by Ladany [13].)

Some efforts were made to estimate the higher order entropies of languages by use of word statistics. Shannon [2], applying the Zipf law, evaluated the entropy of English of approximately 5th or 6th order as 2.62 bits/symbol. Kuepfmueller [12] used the statistics of words and syllables (not including the space as a symbol), but his approach suffers from neglecting the combinations of letters belonging to different neighboring words. He evaluated $G_3 = 2.8$ bits/symbol, $G_6 = 2.0$ bits/symbol and estimated $H \simeq 1.6$ bits/symbol.

The statistical approach is severely restricted by computational difficulties which make us unable to obtain statistics of $L$-grams even for comparatively small $L$ ($L = 10$–$15$).

The use of word statistics and empirical laws of Zipf's law type (the applicability of the latter was very much criticized by Herdan [5], and was rejected by Choueka and Yeshurun [14] with respect to Hebrew) leads to very inaccurate and unreliable results.

Major progress in developing "objective" (statistical) approaches to entropy evaluation for long symbol sequences (not necessarily texts in natural languages) has been achieved by Grassberger [15] who introduced efficient methods based on modifications of the Lempel–Ziv universal coding algorithms. The first applications of these methods to written English showed promising results (the entropy estimates for fiction varied between 1.04 and 1.62 bits per symbol). However, this method shares with other statistical techniques the fundamental limitations which stem from the insufficiency of the total amount of available texts (see below, Section 2.3).

### 2.3. *Shannon's prediction (guessing) method*

Shannon [2] invented an ingenious method which makes it possible to overcome the above-mentioned limitations of the statistical approach and to obtain upper and lower

Table 1. L-gram entropies for eight languages

| $F_L$ | English[1] | Russian[2] | French[3] | Rumanian[4] | Hebrew[5] | Arabic[6] | Portugese[7] | German[8] |
|---|---|---|---|---|---|---|---|---|
| $F_0$ | 4.76 | 5.00 | 4.76 | 4.76 | 4.52[9] | 5.00 | 4.76 | 4.7 |
| $F_1$ | 4.03 | 4.35 | 3.9 | | 4.12[10]<br>4.22[10] | 4.2 | 3.9 | 4.1 |
| $F_2$ | 3.32 | 3.52 | | | 3.71[10] | 3.8 | 3.5 | |
| $F_3$ | 3.1 | 3.0 | 2.83 | 2.69 | 3.21[10] | 2.5 | 3.1 | 2.95 |

[1]Shannon [2].
[2]Garmash *et al.* [45].
[3]Piotrovski [20], Barnard.
[4]Piotrovski [21].
[5]Ladany [47], Ladany [13].
[6]Wanas *et al.* [48].
[7]Manfrino [49].
[8]Piotrovsky [20], Kuepfmueller [12].
[9]23 symbols.
[10]28 symbols.

bounds of the conditional entropy $F_{L+1}(X/S^L)$ for arbitrarily large $L$, so that it gives an evaluation of the entropy per symbol $H(X)$.

The conditional entropy $F_{L+1}$ expresses the uncertainty of a symbol following a set of length $L$ (an $L$-gram). If the entropy is close to zero, it means that the symbol can be predicted almost for certain when the previous $L$-gram is given. The larger the entropy, the more difficult it is to predict the following symbol. This means, in particular, that the predicted symbol can be wrong, and that more than one attempt may be needed to obtain the right result. It suggests to us that it is possible to go in the opposite direction and to extract information about the value of the entropy from the results of a prediction experiment.

Shannon suggested to use a human being—a person experienced in the language—as a predictor for an experiment of this sort. The reasoning is expressed by Shannon in the following words:

The new method of estimating entropy exploits the fact that anyone speaking in a language possesses, implicitly, an enormous knowledge of the statistics of the language. Familiarity with the words, idioms, clichés and grammar enables him to fill in missing or incorrect letters in proof-reading, or to complete an unfinished phrase in conversation.

According to Shannon, the prediction experiment is performed in the following way. The guesser who knows the text up to the current point is asked to guess the next symbol. If he is wrong, he is told so and asked to guess again. This procedure continues until he finds the correct symbol. The number of attempts required until the correct symbol is found is recorded. The next symbol to guess can be either the symbol that follows the guessed one (sequential guessing) or that following a completely different $L$-gram. Here is an example (taken from Shannon's paper) of the results of sequential guessing:

(1) There is no reverse on a motorcycle a
(2) 111511211211 ↓ 1 ↓ 11121321227111141111131
-15-17-
(1) friend of mine found this out
(2) 86131111111111162111111112111111
(1) rather dramatically the other day
(2) 4111111 ↓ 5111111111116111111111111
-11-

The first line is the original text and the numbers in the second line indicate the attempt at which the correct symbol was obtained. The sequence of the recorded numbers of attempts constitutes the so-called "reduced text", which is in fact a specially encoded form of the original text. Shannon has proved that this coding is completely reversible. In the case of non-sequential guessing there is no sense to speak about coding; nevertheless the term "reduced text" will be used.

The numbers of attempts, which are the symbols of the reduced text $y_i$ ($i = 1, 2, \ldots, K$) are in fact natural numbers from 1 to $K$. Because of the randomness of the symbol following a given $L$-gram, the number of attempts is a random variable $Y_L$, which takes on values $y_i$ with probabilities $P_{Y_L}(y_i) = q_i$. These probabilities are determined, of course, by the guesser's performance and can be estimated by the frequencies:

$$\hat{q}_i = \frac{m_i}{N}, \tag{2.3 1}$$

where $N$ is the total number of guessed symbols and $m_i$ is the number of symbols which have been guessed at the $i$th attempt.

Now we shall introduce in a more formal way the random variables $S^L$, $X$ and $Y_L$ with which we are to deal with.

Consider a probability space $(E, \mathcal{B}, P)$, where $E$ is the space of elementary events: $E = \{e_{j,k}\}$, the elementary event $e_{j,k}$ is a pair of a $L$-gram $s_j^L$ and the following symbol $x_k$: $e_{j,k} = (s_j^L, x_k)$, $(j = 1, \ldots, K^L, k = 1, \ldots, K)$; $\mathcal{B}$ is a Borel field on $E$, and $P$ is a probability measure on $E$, defined by the probabilities of the elementary even $P_{S^L,X}(s_j^L, x_k)$. $P_{S^L,X}(s_j^L, x_k)$ is interpreted as a joint probability of the event that $S^L$ takes on a value $s_j^L$ and $X$ takes on a value $x_k$:

$$P_{S^L,X}(s_j^L, x_k) = \Pr[S^L = s_j^L, X = x_k]. \tag{2.3.2}$$

Consider now an element $\mathbf{x}_k$ of the Borel field $\mathcal{B}$ which is the set of all the elementary events $e_{j,k}$ such that $k$ is fixed, and $j$ takes on all the possible values:

$$\mathbf{x}_k = \{e_{j,k} | j = 1, \ldots, K^L\}. \tag{2.3.3}$$

Obviously, $\mathbf{x}_k$ represents the event that the symbol following a $L$-gram is $x_k$. Thus the probability of $X$ to take on the value $x_k$ is

$$P_X(x_k) = \Pr[\mathbf{x}_k] = \Pr[\{e_{j,k} | j = 1, 2, \ldots, K^L\}] = \sum_{j=1}^{K^L} P_{S^L,X}(s_j^L, x_k). \tag{2.3.4}$$

Similarly, if $\mathbf{s}_j^L \in \mathcal{B}$ is the set of all elementary events $e_{j,k}$, such that $j$ is fixed and $k$ takes on all the possible values, then the probability of $S^L$ to take on the value $s_j^L$ is:

$$P_{S^L}(s_j^L) = \Pr[\mathbf{s}_j^L] = \Pr[\{e_{j,k} | k = 1, 2, \ldots, K\}] = \sum_{k=1}^{K} P_{S^L,X}(s_j^L, x_k). \tag{2.3.5}$$

Now let us consider special elements $\mathbf{y}_i$ $(i = 1, 2, \ldots, K)$ of the Borel field $\mathcal{B}$, which are the sets of all the elementary events $e_{j,k}$ such that $j$ takes on all the possible values and for each $j$ and $i$ the index $k$ takes on a value $k = k_i(j)$ in such a way that for different values of $i$ the corresponding values of $k_i(j)$ are also different:

$$\mathbf{y}_i = \{e_{j,k} | j = 1, 2, \ldots, K^L, k = k_i(j)\} \quad (i = 1, 2, \ldots, K; k_{i_2}(j) \neq k_{i_2}(j) \quad \text{if} \quad i_1 \neq i_2). \tag{2.3.6}$$

We interpret the event $\mathbf{y}_i$ as the event that the random variable $Y_L$ takes on the value $y_i$:

$$q_i = P_{Y_L}(y_i) = \Pr(\mathbf{y}_i) = \Pr[\{e_{j,k} | j = 1, \ldots, K^L, k = k_i(j)\}] = \sum_{j=1}^{K^L} P_{S^L,X}(s_j^L, x_{k_i(j)}). \tag{2.3.7}$$

As a matter of fact, the one-to-one correspondence between the indices $i$ and $k_i(j)$ (for a fixed $j$) is a permutation of the alphabetic order of the symbols $x_k$ according to the order in which the symbols are named in the process of guessing a symbol following a given $L$-gram $s_j^L$. It is reasonable to assume that a guesser names symbols consecutively according to his subjective estimation of their conditional probabilities: first the most probable (to his opinion) symbol, then the second one, and so on. This brings us to the concept of an "ideal guesser", or "ideal predictor", introduced by Shannon. An ideal guesser is such a guesser who for any $L$-gram $s_j^L$ knows exactly the order of the conditional probabilities

$$P_{X/X^L}(x_k/s_j^L) = \frac{P_{S^L,X}(s_j^L, x_k)}{P_{S^L}(s_j^L)}, \tag{2.3.8}$$

and names the symbols consecutively in the order of decreasing conditional probability. In other words, for an ideal guesser the indices $k_i(j)$ are chosen in such a way that

$$P_{X/S^L}(x_{k_1(j)}/s_j^L) > (x_{k_2(j)}/s_j^L) \ldots > P_{X/S^L}(x_{k_K(j)}/s_j^L). \tag{2.3.9}$$

Thus for an ideal guesser the probability $q_1$ to guess a letter at the first attempt is the weighted (with probabilities $P_{S^L}(s_j^L)$) sum of all the maximum conditional probabilities, $q_2$ is the weighted sum of the second highest values of conditional probabilities, etc., so that

$$q_1 > q_2 > \ldots > q_K. \tag{2.3.10}$$

The set of probabilities $q_1$ of the numbers of attempts is the only information about the probabilities $P_{S^L,X}(s_j^L, x_k)$, which is obtained in Shannon's guessing experiment. Nevertheless, this limited information allows us to put some upper and lower bounds for the entropy $F_{L+1}$. It was shown by Shannon that for an ideal guesser the following upper and lower bounds for the conditional entropy of $(L+1)$ order are valid:

$$\underline{H}_{\mathrm{Sh}}(X) \underline{\Delta} \sum_{i=1}^{K} (q_i - q_{i+1}) i \ln i \leqslant F_{L+1} \leqslant -\sum_{i=1}^{K} q_i \ln q_i = H(Y_L) \Delta \bar{H}_{\mathrm{Sh}}(X). \tag{2.3.11}$$

(Here and henceforth $A \underline{\Delta} B$ means "$A$ equals $B$ by definition".)

Both of these bounds are attainable; therefore they cannot be improved without any additional information. A real guesser is in fact non-ideal. This means that he misunderstands the order of conditional probabilities of a letter following a $L$-gram at least in some cases. Since $q_i$ is the average of conditional probabilities which occupy the $i$th place in the descending order of the probabilities [cf. (2.3.7)],

$$q_i = \sum_{j=1}^{K^L} P_{S^L}(s_j^L) P_{X/S^L}(x_{k_i(j)}/s_j^L) \tag{2.3.12}$$

a wrong ordering leads to a transfer of some probability from the larger $q_i$ (with smaller index $i$) to the smaller ones (with larger index $i$), so that the distribution of $q_i$ becomes more uniform.

Hence the upper bound increases for a non-ideal guesser, remaining still valid (but less exact). Unfortunately, the lower bound given by (2.3.11) also increases for a non-ideal guesser, which makes the value of the lower bound less reliable. There exists, however, some other effect imposed by the statistical nature of the experimental data which influences in the opposite direction, decreasing the value of the lower bound (cf. Section 4). It should also be taken into account that the lower bound can be achieved only at a very specific (rectangular) form of conditional probability distribution, which is far from actual distributions. Therefore, according to Shannon, the gap between $F_{L+1}$ and the "ideal" lower bound is so wide that it "more than compensates for the failure of human subject to predict in the ideal manner" (Shannon [2]).

Since $F_{L+1} \to H(X)$ when $L \to \infty$, the bounds established by (2.3.11) are also valid for the actual entropy per symbol of the language, if $L$ is large enough. Both the upper and the lower bounds decrease monotonically when $L$ increases, so they do not converge to the entropy $H(X)$. According to some experiments (Burton and Licklider [16]; Piotrovskaia *et al.* [17]) the values of the bounds become practically constant for the length $L$ of the order of several tens (30–100).

The striking power of Shannon's method in comparison with the direct statistical approach becomes evident when we realize that the probabilities of $L$-grams for large $L$ cannot be obtained from statistical data *not only* because of *computational difficulties* but also because the total amount of texts in any language is limited. Indeed, for $L = 50$,

assuming that the entropy per symbol is only 1 bit, we have $2^{50} \cong 10^{17}$ possible different meaningful 50-grams. Let us compare this number with the total length of all the texts published in a given language (say, in English). The library of the USA Congress contains approximately $5 \times 10^7$ volumes. Suppose that, together with all the periodicals, etc., the number of all existing "equivalent volumes" of $10^6$ symbols each is $10^9$. Then the totality of all the texts consists of $10^{15}$ letters. This means that only a small part of all the possible meaningful 50-grams can be found in published texts, and it is very improbable to find the same 50-gram more than once. Thus, the total length of all the existing texts is not enough in order to find probabilities of long $L$-grams. However, a human guesser can usually suggest *several* different continuations of a given possible texts which is much larger than the totality that actually exists. In fact, a human being possesses such an enormous variety of possible texts because of his knowledge of the generating rules of such texts which makes him a source of similar texts himself. This knowledge, which is, of course, mostly intuitive, brings the guesser by reading of a previous text into a state which is close to the state of the source itself, and that enables him to predict efficiently the continuation.

### 2.4. *Sources of errors and limitations of Shannon's method*

In fact Shannon's method does not enable us to calculate the value of $F_L$ itself. Even if the guesser is an ideal one, only upper and lower bounds of $F_L$ can be obtained. It was shown by Shannon [2] that both bounds decrease monotonically with $L$ and approach limits for $L \to \infty$. Nevertheless the limits are, in general, of different values, so there exists a finite gap between the bounds and there is no convergence of the bounds to the actual value of the entropy. For sufficiently large $L$ the difference between $F_L$ and $H(X)$ becomes smaller than the gap between the bounds and therefore the lower bound of $F_L$ becomes also a lower bound of $H(X)$.

The fact that the upper and lower bounds do not converge to the entropy $H(X)$ is a result of rather modest requirements applied to the guesser: he should indicate only the order of conditional probabilities but not their exact values. It was noted by Kolmogorov (see Yaglom and Yaglom [18], p. 257) that if the guesser is able to indicate the values and not only the order of the conditional probabilities of the symbol following a given $L$-gram, then a consistent estimate of entropy per symbol can be constructed, which is given by the arithmetical average of the values of $[-\ln P_{X/S^L}(x_k/s_j^L)$ where $x_k$ is the symbol which actually appears in the text after the given $L$-gram. This approach was implemented in a remarkable paper by Cover and King [19] by use of a gambling procedure: the guesser has to divide a "capital" among all the possible continuations in proportion to values which he assigns to conditional probabilities of various continuations. However, it should be borne in mind that such an approach implies a drastic change in the definition of the ideal guesser. Namely, instead of the order of conditional probabilities, the guesser should know the exact values of them. This requirement seems to be very unrealistic for a human guesser. Even an ideal guesser in Shannon's sense would usually appear to be far from ideal in the sense of Kolmogorov–Cover–King. As shown by Cover and King, an incorrect assessment of the conditional probabilities results always in an overestimation of $F_L$ and of $H(X)$. Thus, in fact, what can be really obtained by use of this approach is not a consistent estimate, but an upper bound of the entropy per symbol of the text; it was also obtained by Shannon's method. The experiment done by Cover and King shows that the upper bound obtained is actually not better than the bound obtained in Shannon's original experiment. But in contrast with Shannon's approach this method gives no opportunity to bound the entropy from below, which is, of course, a serious disadvantage of the method.

Undoubtedly, the most important source of errors in a Shannon–type experiment is the

non-ideality of the guesser. A non-ideal guesser mixes the order of conditional probabilities for some of the $L$-grams, making the distribution $(q_1, q_2, \ldots, q_K)$ less steep. As a result both the upper and the lower bounds become larger. The upper bound then is still valid but becomes more rough. The situation with the lower bound is even worse, because it becomes less reliable: for a guesser far enough from ideality there exists a danger that the calculated value of the lower bound can exceed the entropy.

Another sort of limitation of accuracy is imposed by the statistical nature of the experiment. In fact, the experiment gives us not the probabilities $q_i$ but the frequencies $\hat{q}_i = m_i/N$. The upper and lower bounds of the entropy are to be estimated by use of the frequencies $\hat{q}_i$. It brings into consideration the usual problems inherent in a statistical estimation—those of the variance and the bias of the estimates.

The analysis of the statistical properties of the estimates will be given in Section 5. It should be mentioned that the usual estimate of the upper bound (which is, in fact, an estimate of the entropy of distribution $(q_1, \ldots, q_K)$) given by the formula

$$\hat{\bar{H}}_{\mathrm{Sh}}(X) = \hat{H}(Y_L) = -\sum_{i=1}^{K} \frac{m_i}{N} \ln \frac{m_i}{N} \tag{2.4.1}$$

is considerably negatively biased, which should be taken into account in the evaluation of the upper bound.

The estimate of the lower bound is a linear function of the frequencies:

$$\hat{\underline{H}}_{\mathrm{Sh}}(X) = \sum_{i=1}^{K} \left( \frac{m_i}{N} - \frac{m_{i+1}}{N} \right) i \ln i. \tag{2.4.2}$$

Thus, this estimate is unbiased. But it should be borne in mind that the lower bound in the form (2.3.11) is valid only for monotonically decreasing probability distributions. Therefore the frequencies $\hat{q}_i$ should be smoothed in an appropriate way (see Section 5) in order to satisfy the condition of monotonity. This smoothing leads to a negative bias. On the other hand, this negative bias seems to compensate partially the positive "bias" of the lower bound caused by the non-ideality of the human guesser.

## 2.5. *Early results obtained by Shannon's method*

Shannon [2] was the first to apply the prediction method for the evaluation of the entropy of English. The results were based on a sample of 100 cases of a symbol guessed after a known text of 100 symbols. Numerical values obtained in the experiment are given in Table 2.

Later on, extensive results for a number of European and non-European languages were obtained by Piotrovski [20–22] and his collaborators (Piotrovskaia *et al.* [17]; Petrova *et al.* [23]; Boguslavskaia *et al.* [29, 25]; Novak and Piotrovski [26]; Korolenko *et al.* [27]). Piotrovski introduced an important modification of the guessing method. Namely, he separated the cases when a symbol is uniquely determined by the previous text (so-called "zeros of information"). It gives a considerable improvement of the upper bound of entropy. The results, obtained by Piotrovski *et al.* and by other authors, which used the same method (Baytanaieva and Bektaiev [28]) for fiction samples are given in Table 3. It

Table 2. Results obtained by Shannon for English

| $\hat{\bar{H}}_{\mathrm{Sh}}(X)$ (bits) | $\hat{\underline{H}}_{\mathrm{Sh}}(X)$ (bits) | $\hat{\underline{R}}$ (%) | $\hat{\bar{R}}$ (%) | $\hat{q}_1$ |
|---|---|---|---|---|
| 1.3 | 0.6 | 73 | 87 | 0.8 |

Table 3. Entropy and redundancy of fiction texts for several languages by Piotrovski *et al.*

| | Language | $\hat{\hat{H}}$ bits | $\underline{\hat{H}}$ bits | $\underline{\hat{R}}$ (%) | $\hat{\hat{R}}$ (%) |
|---|---|---|---|---|---|
| 1 | Russian | 1.19 | 0.70 | 76 | 86 |
| 2 | Polish | 1.29 | 0.83 | 74 | 84 |
| 3 | English | 1.10 | 0.65 | 77 | 86 |
| 4 | German | 1.36 | 0.83 | 71 | 82 |
| 5 | French | 1.36 | 0.78 | 71 | 84 |
| 6 | Rumanian | 1.26 | 0.78 | 74 | 84 |
| 7 | Kazakh | 1.35 | 0.81 | 75 | 85 |

should be pointed out that the results given in Table 3 are obtained by sequential guessing as an arithmetical average of $F_L$ for a number of various values of $L$, from $L = 30$ to 100 or 200.

Some other authors used Shannon's original version of the guessing method (Kazarian [29]; Lenskoi [30]; Doležel [31]; Savchuk [32]) (Table 4) or other modifications of the method: a reduced guessing procedure, suggested by Kolmogorov (Piotrovski [20], pp. 60–61) (Table 5); a collective guessing procedure (the latter gives only an estimation of the entropy, which is apparently negatively biased) (Piotrovski [20], pp. 60–61; Petrova *et al.* [23], p. 157; Gut [33]; Korolenko *et al.* [27]; Georgiev [34]) (Table 6). Table 7 contains the results found by Cover and King [19] by use of a "gambling approach", which gives, in fact, an upper bound of entropy.

## 2.6. *Problems involved in the implementation of the guessing method*

The applications of the guessing method usually suffer from a number of shortcomings of different nature which affect the validity and the accuracy of the results.

Table 4. Entropy and redundancy of fiction texts (results modified by Piotrovski by use of a correcting factor)

| | Language | $\hat{\hat{H}}$ bits | $\underline{\hat{H}}$ bits | $\underline{\hat{R}}$ (%) | $\hat{\hat{R}}$ (%) |
|---|---|---|---|---|---|
| 1 | Adyghe | 2.26 | – | 56 | – |
| 2 | Armenian | 1.38 | 0.78 | 74 | 85 |
| 3 | Czech | 1.38 | 1.08 | 74 | 80 |

Table 5. Results obtained by the reduced guessing method Kolmogorov

| | Language | $\hat{H}$ (bits) |
|---|---|---|
| 1 | Russian (fiction) | 1.1 |
| 2 | French | 1.0 |

Table 6. Results obtained by collective guessing

| | Language | $\hat{H}$ (bits) |
|---|---|---|
| 1 | Polish | 0.95 |
| 2 | Rumanian | 0.725 |
| 3 | Bulgarian (fiction) | 0.88 |
| 4 | Spanish | 1.05 |

Table 7. Entropy of English by Cover and King

| | Case description | $\hat{\hat{H}}$ (bits) |
|---|---|---|
| 1 | Best subject estimate (text from "*Jefferson the Virginian*" by D. Malon) | 1.29 |
| 2 | Committee gambling estimate (the same text) | 1.25 |
| 3 | Best subject estimate (text from "*Contact*" by L. and N. Zunin) | 1.26 |

First of all, it is obvious that in the case of sequential guessing many of the $L$-grams used in the experiment are largely overlapping and thus cannot be considered as independent samples. It affects the representativeness of the sampling and makes the statistical treatment of the results impossible. Therefore, in our opinion, a non-sequential guessing procedure which uses non-overlapping $L$-grams chosen from different and remote parts of the text is much more preferable.

It should be noted that the uncertainty of a letter to be guessed depends very much on the position of the letter in a word. For instance, it is minimum at the end of the word and maximum at the beginning of it. Therefore, if the text given to a guesser in a sequential guessing experiment starts with the beginning of a word (as it was usually done) the results display a quasi-periodic oscillatory structure due to the fact that letters situated at different distances from the beginning of the text have different probabilities to occupy a certain position in a word. The experiments made by Piotrovski and others show that even for lengths of several hundreds of letters and text still keeps memory about the beginning. In order to eliminate this memory effect and to obtain valid bounds of $F_{L+1}$ for a given $L$, the beginning of each sampled $L$-gram should be chosen completely at random (cf. Section 7).

In the original version of Shannon's method the probability $q_i$ estimated from an experiment is a result of averaging conditional probabilities $P_{X/S^L}(x_{k_i(j)}/s_j^L)$ for fixed $i$ over the set of all linguistic situations (cf. (2.3.12)). Hence, the conditional entropy $F_{L+1}$, i.e. the averaged value of the entropy of conditional probability distributions (cf. (2.1.1)), is replaced by the entropy of averaged conditional probabilities $q_i$, which gives an upper bound of the former. Since entropy is a concave function of probability, the more the conditional probabilities $P_{X/S^L}(x_{k_i(j)}/s_j^L)$ for a given $i$ differ for different $j$s, the more the upper bound differs from the actual value of the conditional entropy. In fact, the values of the above-mentioned conditional probabilities are very different for different linguistic situations (for instance, the probability of the most probable continuation can vary from $1/K$ up to 1). Thus the upper bound given by the original Shannon method is rather rough.

An attempt to overcome this disadvantage was undertaken by Piotrovskaia *et al.* [17]. They separated linguistic situations when the following letter is determined uniquely by the previous $L$-gram. The separation of the "zeros of information" implies the use of some objective characteristics of the text and can be performed even a posteriori, after the termination of the guessing experiment. Another possible approach is to extract some additional information from the guesser. In fact, a good guesser can tell more about the probabilities of possible continuations than just their order, as is assumed for an ideal guesser by Shannon.

Kolmogorov and his co-workers performed a number of experiments aimed at obtaining such information (Rychkova [6]). They suggested to the guesser to make one of the following predictions:

(1) the next symbol is certainly the $k$th letter of the alphabet;
(2) the next symbol is one of two or three letters named by the guesser;
(3) the next symbol is probably (but not certainly) the $k$th letter of the alphabet;
(4) the next letter is probably one of two or three letters; or
(5) no prediction can be made by the guesser.

This approach yielded an upper bound of 1.0–1.2 bits per symbol (for Russian language).

Another extreme assumption is that an ideal guesser is able to evaluate exactly the conditional probabilities of all the possible continuations after a given $L$-gram (Cover and King [19]). The best strategy is, probably, to put before a human guesser such questions which are adequate to his ability to discriminate probabilities of various possible continuations.

The situation with the lower bound is even worse. As was pointed out by Shannon [2] his lower bound is far from the actual value of $F_{L+1}$ due to "the failure to have rectangular distribution of conditional probability". To the best of our knowledge, no research was done previously aimed at improving the lower bound given by Shannon.

There are a number of statistical problems involved in the evaluation of the bounds of entropy (see Section 5). Seemingly, no statistical treatment was given before to the results of prediction experiments.

## 3. AN IMPROVED METHOD FOR THE DETERMINATION OF THE UPPER BOUND OF ENTROPY

### 3.1. *Subdivision of the set of* L-*grams*

According to Shannon's prediction method the only information about the probability distribution $P_{S^L,X}(s_j^L, x_k)$ obtained from the guessing experiment is that contained in the probabilities $q_i = P_{Y_L}(y_i)$, which are, in essence, the result of averaging of the ordered conditional probabilities of a letter $x_{k_i(j)}$ following a sequence $s_j^L$:

$$q_i = P_{Y_L}(y^i) = \sum_{s_j^L} P_{S^L}(s_j^L) P_{X/S^L}(x_{k_i(j)}/s_j^L). \tag{3.1.1}$$

The roughness of the upper bound given by

$$\bar{H}_{\mathrm{Sh}}(X) = H(Y_L) = -\sum_i q_i \ln q_i \tag{3.1.2}$$

depends on the dispersion of values of the conditional probabilities $P_{X/S^L}(x_{k_i(j)}/s_j^L)$ for given $i$. The more they differ the greater the increase of entropy resulting from the smoothing given by (3.1.1). This suggests improving the upper bound by subdivision of the totality $\Omega$ of all possible sequences $s_j^L$ into subsets which are more homogeneous with respect to the conditional probability distributions $P_{X/S^L}(x_{k_i(j)}/s_j^L)$.

Consider the simplex of all probability distributions of ordered probabilities $p_i = P_{X/S^L}(x_{k_i(j)}/s_j^L)$ defined by inequalities:

$$p_1 \geqslant p_2 \geqslant p_3 \geqslant \ldots \geqslant p_K \geqslant 0 \qquad \text{and} \qquad \sum_{i=1}^{K} p_i = 1. \tag{3.1.3}$$

Our aim is to subdivide the simplex into domains as small as possible but still distinguishable by the guesser. From our experience a human guesser can estimate well only the values of the largest probabilities and their sums: $p_1$, $p_1 + p_2$, ... up to $p_1 + p_2 + p_3 + p_4$. This is the reason for defining the subdivision in terms of some threshold values of the sums of the largest probabilities.

Some other a priori characteristics can also indicate domains in the probability simplex in an indirect way and so they can be used for the subdivision of the set of all $L$-grams.

3.2. *The effect of subdivision on the upper bound of the entropy*

Consider the set $\Omega$ of all possible sequences $s_j^L$ of length $L$ and a partition of $\Omega$ into a number of disjoint subsets $a_r$. Let us introduce the average conditional probabilities $q_i^{(r)}$ of a symbol following a sequence $s_j^L$ belonging to a given subset $a_r$:

$$q_i^{(r)} = \frac{\sum\limits_{s_j^L \in a_r} P_{S_L,X}(s_j^L, x_{k_i(j)})}{\sum\limits_{s_j^L \in a_r} P_{S^L}(s_j^L)}, \tag{3.2.1}$$

where $q_i^{(r)}$ is the probability of a symbol to be guessed at the $i$th attempt if the preceding $L$-gram $s_j^L$ belongs to $a_r$. Denote

$$\omega_r = \sum_{s_j^L \in a_r} P_{S^L}(s_j^L). \tag{3.2.2}$$

Then Shannon's upper bound of the entropy per symbol for the cases when $s_j^L \in a_r$ is given by

$$\bar{H}_r = -\sum_i q_i^{(r)} \ln q_i^{(r)}. \tag{3.2.3}$$

Consider the weighted sum of the upper bounds $\bar{H}_r$ for the subsets $a_r$:

$$\bar{H}(X) \Delta \sum_r \omega_r \bar{H}_r = -\sum_r \sum_i \omega_r q_i^{(r)} \ln q_i^{(r)}. \tag{3.2.4}$$

The next two theorems show that the quantity $\bar{H}(X)$ gives a better upper bound of the entropy $F_{L+1}$ than the quantity $\bar{H}_{\mathrm{Sh}}(X)$ given by Shannon, i.e., that any subdivision of $\Omega$ can only improve the upper bound.

*Theorem* 3.2.1.

$$F_{L+1}(X/S^L) \leqslant \bar{H}(X). \tag{3.2.5}$$

The equality holds only in the case when $q_i^{(r)} = P_{X/S^L}(x_{k_i(j)}/s_i^{(r)})$ for any $s_j^L \in a_r$.

*Theorem* 3.2.2.

$$\bar{H}(X) \leqslant \bar{H}_{\mathrm{Sh}}(X), \tag{3.2.6}$$

where the equality holds iff the probabilities $q_i^{(r)}$ are the same for all the subsets $a_r$ and do not depend on $r$.

## 4. AN IMPROVED METHOD FOR THE DETERMINATION OF THE LOWER BOUND OF ENTROPY

4.1. *A simplex of ordered probabilities*

As was mentioned above, for any sequences $s_j^L$ the ordered conditional probabilities $p_i = P_{X/S^L}(x_{k_i(j)}/s_j^L)$, $(i = 1, 2, \ldots, K)$, can be considered as components of $K$-dimensional vector belonging to a $(K-1)$-dimensional simplex. More exactly, the following proposition is valid:

*Theorem* 4.1.1. The set of all possible vectors of ordered probabilities $\mathbf{p} = (p_1, \ldots, p_K)$ constitutues a $(K-1)$-dimensional simplex $\mathbb{O}$ in a $K$-dimensional linear space defined by conditions:

$$p_1 \geqslant p_2 \geqslant \ldots \geqslant p_K \geqslant 0; \qquad \sum_{i=1}^{K} p_i = 1 \tag{4.1.1}$$

with $K$ vertices:

$$\mathbf{v}_g = \left(\frac{1}{g}, \ldots \frac{1}{g}, 0, \ldots 0\right); \qquad (g = 1, \ldots, K) \tag{4.1.2}$$

where the vector $\mathbf{v}_g$ has $g$ non-zero components.

## 4.2. *Entropy as a concave function on a convex set; Shannon's lower bound*

It is known (Fano [35]) that the entropy of a discrete random variable $Z$ taking values $z_i$ with probabilities $p_i$ $(i = 1, \ldots, K)$ is a concave function of the probability vector $\mathbf{p} = (p_1, \ldots, p_K)$. That is, for a number of probability vectors $\mathbf{p_j} = (p_{j1}, \ldots, p_{jK})$ and any weight coefficients $\alpha_j \geqslant 0$, $\sum_j \alpha_j = 1$, the following inequality is valid:

$$\sum_j \alpha_j H(\mathbf{p_j}) \leqslant H\left(\sum_j \alpha_j \mathbf{p_j}\right) \tag{4.2.1}$$

where $H(\mathbf{p_j}) = -\sum_i p_{ji} \ln p_{ji}$, and the equality holds iff all the probability vectors $\mathbf{p_j}$ are identical.

Consider the set of probabiity vectors $\mathbf{p_j} = (p_{j1}, \ldots, p_{jK})$ of ordered conditional probabilities $p_{ij} = P_{X/S^L}(x_{k_i(j)}/s_j^L)$. These conditional probabilities are unknown, but the prediction method allows us to evaluate the mean probability vector

$$\mathbf{q} = (q_1, \ldots, q_K) = \sum_j P_{S^L}(s_j^L) p_j. \tag{4.2.2}$$

The problem is to find a lower bound for the conditional entropy

$$F_{L+1}[X/S^L] = \sum_j P_{S^L}(s_{jL}) H(\mathbf{p_j}) \tag{4.2.3}$$

based on the knowledge of the mean probability vector $\mathbf{q}$ only.

A non-trivial lower bound can be found in the case when all the probability vectors $\mathbf{p_j}$ belong to a convex set $C$, which is a convex hull of its extreme points $\mathbf{u_g}$ $(g = 1, \ldots, G, G \geqslant K)$.

*Theorem* 4.2.1. Let all $\mathbf{p_j} \in C$ and $\mathbf{q} = \sum_j P_{S^L}(s_j^L)\mathbf{p_j}$, then the conditional entropy $F_{L+1}(X/S^L)$ satisfies inequality

$$F(X/S^L) \geqslant \min_{\{\alpha_g\}} \sum_{g=1}^{G} \alpha_g H(\mathbf{u}_g), \tag{4.2.4}$$

where the minimum is taken over all the possible sets of weight coefficients $\{\alpha_g\}$ subjected to constraints:

$$\alpha_g \geqslant 0; \qquad \sum_{g=1}^{G} \alpha_g = 1; \qquad \sum_{g=1}^{G} \alpha_g \mathbf{u}_g = \mathbf{q}. \tag{4.2.5}$$

It should be borne in mind that the condition $\sum_{g=1}^{G}\alpha_g = 1$ is not independent, but follows from the equation $\sum_{g=1}^{G}\alpha_g \mathbf{u}_g = \mathbf{q}$ and from the fact that $\mathbf{q}$ and all $\mathbf{u}_g$ are probability vectors.

The equality in (4.2.4) holds iff each of the probability vectors $\mathbf{p_j}$ coincides with one of vertices $\mathbf{u}_g$.

Note also that the sum in the right-hand side of (4.2.4) is a linear function of the weights $\alpha_g$ over a compact set $C$, and therefore the minimum is always attainable (Bourbaki [36]).

In our case, as was shown in Section 4.1, all the vectors $\mathbf{p_j}$ belong to the simplex $\mathbb{O}$, defined by (4.1.1). For any set of such vectors the following theorem is valid.

*Theorem* 4.2.2. Let all $\mathbf{p_j} \in \mathbb{O}$ and $\mathbf{q} = (q_1, \ldots, q_g, \ldots, q_K) = \sum_j P_S(s_j^L)\mathbf{p_j}$.
Then the lower bound for the conditional entropy $F(X/S^L)$ is given by inequality

$$F_{L+1}(X/S^L) \geqslant \sum_{g=1}^{K}(q_g - q_{g+1})g \ln g. \tag{4.2.6}$$

The lower bound (4.2.6) was given first by Shannon [2], who used a different way of reasoning. This bound is attained iff every $\mathbf{p_j}$ coincides with one of the vertices $\mathbf{v}_g$: $\mathbf{p_j} = \mathbf{v}_{g(j)}$.

## 4.3. *Subdivision of the simplex $\mathbb{O}$ into convex domains; its effect on the lower bound of entropy*

The lower bound (4.2.6), being attainable, cannot be improved, if the only information about the conditional probability vectors $p_j$ is the knowledge of the mean probability vector $\mathbf{q}$. But if, in addition, the set $\Omega$ of all sequences $s_j^L$ is subdivided into disjoint subsets $d_r$, $\bigcup_r d_r = \Omega$, such that if $s_j^L \in d_r$ then $\mathbf{p_j} \in C_r$, where $C_r \subseteq \mathcal{o}$ is a convex subset of the simplex $\mathbb{O}$, and for each subset $d_r$ the mean probability vector

$$\mathbf{q^{(r)}} = \frac{\sum_{s_j^L \in d_r} P_S(s_j^L)\mathbf{p_j}}{\sum_{s_j^L \in \mathbf{d_r}} P_S(s_j^L)} \tag{4.3.1}$$

is given, this additional information can be used in order to obtain a better lower bound for $F_{L+1}(X/S^L)$.

Consider a number of convex sets $C_r \subseteq \mathcal{o}$, $\bigcup_r C_r = \mathcal{o}$ with extreme points $\mathbf{u}_{g_r}^{(r)}$, ($g_r = 1, \ldots, G_r$; $G_r \geqslant K$). (Sets $C_r$ should not be, in general, disjoint.) Denote the probability that $s_j^L \in d_r$ by $\rho_r$

$$\rho_r = \sum_{s_j^L \in d_r} P_S(s_j^L).$$

(Note, that if $s_j^L \in d_r$, then $\mathbf{p_j} \in C_r$, but the converse is, in general, not correct.) Then it follows from Theorem 4.2.1 applied to each subset $d_r$, that

$$F_{L+1}(X/S^L) \geqslant \sum_r \rho_r \min_{\{\alpha_{g_r}^{(r)}\}} \sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} H(\mathbf{u}_g^{(r)}), \tag{4.3.2}$$

where

$$\alpha_{g_r}^{(r)} \geqslant 0; \qquad \sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} u_{g_r}^{(r)} = q^{(r)}. \tag{4.3.3}$$

The next theorem shows that the lower bound given by (4.3.2) is better than the lower bound given by (4.2.6).

*Theorem* 4.3.1. Consider a convex $(K-1)$-dimensional set $C$ with extreme points $\mathbf{u}_g$ $(g = 1, \ldots, G,\ G \geqslant K)$ in a $K$-dimensional linear space. Assume that the convex set $C$ corresponds to a set $d \subseteq \Omega$ of sequences $s_j^L$, so that if $s_j^L \in d$ then $\mathbf{p_j} \in C$.

Denote:

$$\rho = \Pr\{s_j^L \in d\} = \sum_{s_j^L \in d} P_{S^L}(s_j^L). \tag{4.3.4}$$

Let $C_r$ $(r = 1, 2, \ldots)$ be convex subsets $C_r \subseteq C$, $\bigcup_r C_r = C$ with extreme points $\mathbf{u}_{g_r}^{(r)}(g_r = 1, \ldots, G_r;\ G_r \geqslant K)$. ($C_r$ can intersect one another and some of the extreme points $\mathbf{u}_{g_r}^{(r)}$ can coincide with $\mathbf{u}_g$ and can be common to several different subsets.)

Let $d_r$ be disjoint subsets of $d$, $\bigcup_r d_r = d$.

Denote:

$$\rho_r = \sum_{s_j^L \in d_r} P_{S^L}(s_j^L), \qquad \sum_r \rho_r = \rho; \tag{4.3.5}$$

$$\mathbf{q}^{(r)} = \frac{1}{\rho_r} \sum_{s_j^L \in d_r} P_{S^L}(s_j^L)\mathbf{p_j}; \tag{4.3.6}$$

$$\mathbf{q} = \frac{1}{\rho} \sum_{s_j^L \in d} P_{S^L}(s_j^L)\mathbf{p_j} = \frac{1}{\rho} \sum_r \rho_r \mathbf{q}^{(r)}. \tag{4.3.7}$$

Then the following inequality is valid:

$$\sum_r \rho_r \min_{\{\alpha_{g_r}^{(r)}\}} \sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} H(\mathbf{u}_{g_r}^{(r)}) \geqslant \rho \min_{\{\alpha_g\}} \sum_{g=1}^{G} \alpha_g H(\mathbf{u}_g), \tag{4.3.8}$$

where:

$$\alpha_{g_r}^{(r)} \geqslant 0; \qquad \sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} \mathbf{u}_{g_r}^{(r)} = \mathbf{q}^{(r)} \tag{4.3.9}$$

$$\alpha_g \geqslant 0; \qquad \sum_{g=1}^{G} \alpha_g \mathbf{u}_g = \mathbf{q}. \tag{4.3.10}$$

The inequality (4.3.8) turns into equality iff several very specific requirements are fulfilled:

1. No new vertices are introduced by subdivision: each of $\mathbf{u}_{g_r}^{(r)}$ coincides with one of $\mathbf{u}_g$.
2. There exists a basis (a set of $K$ vertices) common to all the convex domains $C_r$.
3. All the vectors $\mathbf{q}^{(r)}$ belong to the convex hull of the common basis of requirement 2.
4. This common basis is that for which the minimum of the quantity $\sum_{g=1}^{G_r} \alpha_{g_r}^{(r)} H(\mathbf{u}_{g_r}^{(r)})$ is attained simultaneously for all $r$.

Theorem 4.3.1 shows, in particular, that the lower bound of entropy, as defined by (4.2.4), is a convex function of the mean probability vector $\mathbf{q}$.

Applying this theorem to the simplex $\mathbb{C}$, we obtain an affirmative answer to the question which was formulated at the beginning of this section.

Thus, the improved lower bound is given by

$$\underline{H}(X) \Delta \sum_r \min_{\{\alpha_{g_r}^{(r)}\}} \sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} H(\mathbf{u}_{g_r}^{(r)}) \tag{4.3.11}$$

where

$$\alpha_{g_r}^{(r)} \geqslant 0; \qquad \sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} \mathbf{u}_{g_r}^{(r)} = \mathbf{q}^{(r)}. \tag{4.3.12}$$

### 4.4. *Determination of the lower entropy bound as a problem of linear programming*

The results of the preceding section show that if we can use some additional a priori data in order to implement the subdivision of the set $\Omega$ of all sequences into disjoint sets $d_r$ and of the simplex $\mathbb{O}$ into corresponding convex subsets $C_r$, an improvement of the lower bound is to be expected. Now a new problem arises, namely, to find for each convex set $C_r$ the minimum of the expression

$$\sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} H(\mathbf{u}_{g_r}^{(r)}) = F \tag{4.4.1}$$

where $\alpha_{g_r}^{(r)}$ are unknown variables, but subjected to the conditions (4.3.12).

If the minimum of (4.4.1) is obtained for $\alpha_{g_r}^{(r)} = \beta_{g_r}^{(r)}$, then the lower bound of entropy for the subset $d_r$ is given by:

$$\underline{H}_r = \min_{\{\alpha_{g_r}^{(r)}\}} \sum_{g_r=1}^{G_r} \alpha_{g_r}^{(r)} H(\mathbf{u}_{g_r}^{(r)}) = \sum_{g_r=1}^{G_r} \beta_{g_r}^{(r)} H(\mathbf{u}_{g_r}^{(r)}). \tag{4.4.2}$$

Note, that $F$ is a linear function of $\alpha_{g_r}^{(r)}$, while $H(\mathbf{u}_{g_r}^{(r)})$ are given numbers. In the case $G_r = K$, equations (4.3.12) have a unique solution, and the value of $\underline{H}_r$ is uniquely determined. But in case $G_r > K$ there exist an infinite variety of coefficient sets $\alpha_{g_r}^{(r)}$ satisfying conditions (4.3.12). Then we come to a classical linear programming problem, namely, to find the minimum of a linear function (4.4.1) under a system of linear constraints (4.3.12). As is known from the theory of linear programming (Danzig [37]), the minimum is always achieved at an extreme point of the convex set which is the set of all possible solutions of (4.3.12). The dimensionality of this convex set is $G_r$, and its extreme points are characterized by the condition, that $(G_r - K)$ of the weights $\alpha_{g_r}^{(r)}$ are equal to zero, while the rest $K$ weights are determined uniquely as a solution of the system (4.3.12). (The inequalities $\alpha_{g_r}^{(r)} \geqslant 0$ are satisfied automatically for $g^{(r)} \in C_r$.) If all the $K$ weights are actually non-zero, then the number of the extreme points is equal to

$$\binom{G_r}{K} = \frac{G_r!}{K!(G_r - K)!}.$$

Of course, even for moderately large $G_r$ the selection of the extreme point which yields the minimum requires the use of a computer.

## 5. DETERMINATION OF THE UPPER AND LOWER BOUNDS OF ENTROPY AS A STATISTICAL PROBLEM

### 5.1. *The estimate of the upper bound and its properties*

The upper bound of the entropy per symbol obtained by subdivision of the set of $\Omega$ of all the possible sequences $s_j^L$ into disjoint subsets $a_r$ is the mean value of the upper bounds for each of the subsets $a_r$ (see (3.2.3), (3.2.4)).

In fact, the data obtained from an experiment are not the probabilities of the subsets $\omega_r$

and the mean conditional probabilities $q_i^{(r)}$, but the numbers of occurences $N(a_r) = N_r$ and $m_i(a_r) = m_i^{(r)}$. The maximum likelihood estimates of probabilities are given by frequencies

$$\hat{\omega}_r = \frac{N_r}{N} \tag{5.1.1}$$

$$q_i^{(r)} = \frac{m_i^{(r)}}{N_r}. \tag{5.1.2}$$

Here of course, $\sum_{i=1}^{K} m_i^{(r)} = N_r$, $\sum_{r=1}^{R} N_r = N$, where $R$ is the number of subsets $a_r$. Then the simplest estimate of the upper bound can be obtained by substituting the frequencies for the probabilities in (5.1.1) and (5.1.2):

$$\hat{\bar{H}}(X) = \sum \frac{N_r}{N} \hat{\bar{H}}_r \tag{5.1.3}$$

$$\hat{\bar{H}}_r = -\sum_{i=1}^{K} \frac{m_i^{(r)}}{N_r} \ln \frac{m_i^{(r)}}{N_r}. \tag{5.1.4}$$

The numbers of occurrences $N_r$ have a multinomial distribution:

$$\Pr(N_1, \ldots, N_r) = N! \prod_{r=1}^{R} \frac{1}{N_r!} \omega_r^{N_r}. \tag{5.1.5}$$

The estimates $\hat{\bar{H}}_r$ of the upper bounds for the subsets $a_r$ are independent random variables. The joint probability distribution of all $N_r$ and $H_r$ can be written in a form:

$$\Pr(N_1 \ldots, N_R, \hat{\bar{H}}_1, \ldots, \hat{\bar{H}}_R) = \Pr(N_1, \ldots, N_R) \prod_{r=1}^{R} \Pr(\hat{\bar{H}}_r). \tag{5.1.6}$$

Using (5.1.5) and (5.1.6) we obtain the following expression for the variance of the estimate $\hat{\bar{H}}(X)$:

$$V(\hat{\bar{H}}(X)) = \sum_r \omega_r^2 V(\hat{\bar{H}}_r) + \frac{1}{N}\left\{\sum_r \omega_r(1 - \omega_r) V(\hat{\bar{H}}_r) \sum_r \omega_r E^2(\hat{\bar{H}}_r) - E^2\left(\sum_r \omega_r \hat{\bar{H}}_r\right)\right\}, \tag{5.1.7}$$

where $V(Z)$ and $E(Z)$ denote the variance and the expectation of a random variable $Z$, respectively.

It is seen from (5.1.7) that the main term of the total variance is a linear combination of the variances of the estimates for separate subsets. This term would be the only one, had the numbers $N_r$ of cases belonging to each subset not been random. The secondary term (of order $1/N$) appears due to the random sampling of the $L$-grams. It is easy to show that this term is always positive.

For the bias of the estimate $\hat{\bar{H}}(X)$ using (5.1.5) and (5.1.6) we obtain:

$$B(\hat{\bar{H}}(X)) = E(\hat{\bar{H}}(X) - \bar{H}(X)) = E\left(\sum_r \frac{N_r}{N} \hat{\bar{H}}_r\right) - \bar{H}(X)$$

$$= \sum_r \omega_r E(\hat{\bar{H}}_r) - \sum_r \omega_r \bar{H}_r = \sum_r \omega_r [E(\hat{\bar{H}}_r) - \bar{H}_r]$$

$$= \sum_r \omega_r B(\hat{\bar{H}}_r) \tag{5.1.8}$$

According to formulae (5.1.7) and (5.1.8) the value of the total variance is determined mainly by a weighted sum of variances of estimates $\hat{\bar{H}}_r$ and the total bias is just the mean value of the biases of the same estimates. Each $\hat{\bar{H}}_r$ is an estimate of the entropy of a discrete random variable evaluated from a sequence of independent trials. The properties of the estimates have been studied by Basharin [38], and Levitin and Reingold [39]. The results show that the variance of the estimate for each group can be written in a form (omitting terms of order $1/N_r^2$)

$$V(\hat{\bar{H}}_r) = \frac{1}{N_r}\left[\sum_{i=1}^{K} q_i^{(r)} \ln^2 q_i(r) - \hat{H}_r^2\right], \tag{5.1.9}$$

and the bias (omitting terms of order $1/N_r^3$) for the case $q_i^{(r)} N_r \gg 1$ is

$$B(\hat{\bar{H}}_r) = -\frac{K-1}{2N_r} - \sum_{i=1}^{K} \frac{1 - q_i^{(r)2}}{12 q_i^{(r)2} N_r^2}. \tag{5.1.10}$$

Practically, in our case, $K = 23$ and $N_r$ is of the order of 100. Thus the bias is rather large and cannot be neglected. For practical use we have to substitute the estimates $\hat{\bar{H}}_r$ for their expected values $E(\hat{\bar{H}}_r)$ and for the upper bounds of entropy $\bar{H}_r$ and replace the probabilities $q_i^{(r)}$ by frequencies $\hat{q}_i^{(r)} = m_i^{(r)}/N_r$, and probabilities $\omega_r$ by frequencies $\hat{\omega}_r = N_r/N$.

The result is an estimate of the variance of the estimate of the upper bound of entropy per symbol.

$$\hat{V}(\hat{\bar{H}}(X)) = \frac{1}{N^2}\sum_{r=1}^{R} N_r^2 \hat{V}(\hat{\bar{H}}_r) + \frac{1}{N}\left\{\frac{1}{N^2}\sum_{r=1}^{R} N_r(N - N_r)\hat{V}(\hat{\bar{H}}_r) + \frac{1}{N}\sum_{r=1}^{R} N_r \hat{H}_r^2 - \hat{\bar{H}}^2(X)\right\} \tag{5.1.11}$$

where

$$\hat{V}(\hat{\bar{H}}_r) = \frac{1}{N_r}\left[\sum_{i=1}^{K} \frac{m_i^{(r)}}{N_r} \ln^2 \frac{m_i^{(r)}}{N_r} - \hat{\bar{H}}_r^2\right]. \tag{5.1.12}$$

The estimate of the bias is given by

$$\hat{B}(\hat{\bar{H}}(X)) = \sum_{r=1}^{R} \frac{N_r}{N} \hat{B}(\hat{\bar{H}}_r) \tag{5.1.13}$$

where $\hat{B}(\hat{\bar{H}}_r)$ is evaluated numerically.

## 5.2. *Estimation of the lower bound*

As it was shown in Section 4.4, the value of $\underline{H}(X)$ is determined by a solution of a linear programming problem (4.3.11), (4.3.12). For every set of values $q_i^{(r)}$ $(i = 1, \ldots, K)$ not more than $K$ coefficients $\alpha_{g_r}^{(r)}$ are nonzero. It is known from the theory of linear programming that $\alpha_{g_r}^{(r)}$ are piece-wise linear functions of $q_i^{(r)}$. It means that, in general, there exists a neighbourhood of a probability vector $\mathbf{q}^{(r)}$, where the lower bound is a linear function of the probabilities $q_i(r)$.

In order to estimate $\underline{H}(X)$ from the experimental data we have to use frequencies

$$\hat{\rho}_r = \frac{N_r}{N}, \qquad \hat{q}_i^{(r)} = \frac{m_i^{(r)}}{N_r} \tag{5.2.1}$$

instead of the probabilities $\rho_r$ and $q_i^{(r)}$.

However, here we meet a difficulty resulting from the fact that the estimate $\mathbf{q}^{(r)}$ does not necessarily belong to the convex domain $C_r$, because of statistical fluctuations of the number of occurrences $m_i^{(r)}$ and because of the non-ideality of the guesser. In such a case the conditions (4.3.12) become incompatible. It means that in such cases not the frequencies $\hat{q}_i^{(r)}$ but some other estimates of the probabilities $q_i^{(r)}$ should be used which always satisfy the a priori order restrictions.

### 5.3. *The maximum likelihood of probabilities with known order*

Consider a discrete random variable $Y$ which takes on values $y_i$ $(i = 1, \ldots, K)$ with probabilities $q_i$. Suppose the probabilities $q_i$ are unknown, but it is known a priori that the values of the probabilities are ordered:

$$q_1 \geqslant q_2 \geqslant \ldots \geqslant q_K \geqslant 0; \qquad \sum_{i=1}^{K} q_i = 1. \tag{5.3.1}$$

Now suppose that the results of $N$ independent trials are known, the random variable $Y$ taking on the value $y_i$ in $m_i$ trials. $\sum_{i=1}^{K} m_i = N$. It is known that in the case where there are no special restrictions on the values of the probabilities $q_i$ the maximum likelihood estimation of the probabilities is given by the frequencies $q_i^* = m_i/N$ which provide the absolute maximum value of the likelihood function for the multinominal distribution (e.g. Van der Waerden [40]):

$$L(\mathbf{m}/\mathbf{q}) = L(m_1, \ldots, m_K/q_1, \ldots, q_K) = \prod_{i=1}^{K} q_i^{m_i}. \tag{5.3.2}$$

However, the frequencies $q_i^*$ do not satisfy, in general, conditions (5.3.1). Therefore, we have to find the constrained maximum of the likelihood function (5.3.2) subjected to conditions (5.3.1). The solution of the problem is given by the following propositions.

*Lemma* 5.3.1. If for some $i$,

$$m_i < m_{i+1} \tag{5.3.3}$$

then the likelihood function $L$ can achieve maximum only on the boundary of the simplex $\mathbb{O}$ defined by the condition:

$$q_i = q_{i+1}. \tag{5.3.4}$$

*Lemma* 5.3.2. If the likelihood function $L(\mathbf{m}/\mathbf{q})$ can achieve maximum value only on the boundary of the simplex $\mathbb{O}$, defined by conditions:

$$q_i = q_{i+1} = \ldots = q_{i+k-1}$$

and

$$q_{i+k} = q_{i+k+1} = \ldots = q_{i+k+t-1} \tag{5.3.5}$$

and at the same time

$$\frac{m_i + \ldots + \ldots + m_{i+k-1}}{k} < \frac{m_{i+k} + \ldots + m_{i+k+t-1}}{t}, \tag{5.3.6}$$

then the likelihood function $L(\mathbf{m}/\mathbf{g})$ can achieve maximum value only on the boundary

defined by the condition

$$q_i = q_{i+1} = \ldots = q_{i+k} = \ldots = q_{i+k+t-1}. \tag{5.3.7}$$

*Theorem* 5.3.1. If the likelihood function $L$ can achieve maximum value only on the boundary of the simplex $\mathbb{O}$ defined by conditions:

$$q_{s_r+1} = q_{s_r+2} = \ldots = q_{s_{r+1}} \tag{5.3.8}$$

where

$$r = 0, 1 \ldots n-1, \qquad 0 < s_0 < s_1 < \ldots < s_n = K \tag{5.3.9}$$

and for any $r$:

$$\frac{m_{s_r+1} + \ldots + m_{s_{r+1}}}{s_{r+1} - s_r} \geqslant \frac{m_{s_{r+1}+1} + \ldots + m_{s_{r+2}}}{s_{r+2} - s_{r+1}} \tag{5.3.10}$$

then the maximum likelihood estimate for the probabilities $q_1, \ldots, q_K$ are given by:

$$\breve{q}_{s_r+1} = \ldots = \breve{q}_{s_{r+1}} = \frac{m_{s_r+1} + \ldots + m_{s_{r+1}}}{N(s_{r+1} - s_r)}. \tag{5.3.11}$$

## 5.4. *The estimate of the lower bound and its properties*

Using the results of Section 5.3, we can now write the estimate of the lower bound of the entropy in the form

$$\underline{\hat{H}}(X) = \sum_r \sum_{g_r=1}^{G_r} \hat{p}_r \beta_{g_r}^{(r)} H(\mathbf{u}_{g_r}^{(r)}), \tag{5.4.1}$$

where $\beta_{g_r}^{(r)}$ are the coefficients for which the right-hand side of (5.4.1) achieves maximum under conditions:

$$\beta_{g_r}^{(r)} \geqslant 0; \qquad \sum_{g_r=1}^{G_r} \beta_{g_r}^{(r)} = \mathbf{q}^{(r)}. \tag{5.4.2}$$

The "smoothing" of the frequencies according to the rules derived in Section 5.3 always makes the distribution $\breve{q}_i^{(r)}$ steeper than $\hat{q}_i^{(r)}$, which results in a lower value of $\underline{\hat{H}}(X)$. Therefore, $\underline{\hat{H}}(X)$ obtains a negative bias. On the other hand, $\underline{\hat{H}}(X)$ is a convex function of the vector $\mathbf{q}^{(r)}$ (cf. Theorem 4.3.1). Therefore, the expected value of the estimate of the lower bound is larger than the lower bound calculated by use of the expected values of the smoothed frequencies. (These expected values constitute a distribution which is steeper than the actual probability distribution $q_i^{(r)}$.) This means that the estimate $\underline{\hat{H}}(X)$ gets a positive bias, which is opposite to the bias caused by the smoothing of the frequencies. Due to the compensation of these two influences the total bias of the estimate of a lower bound seems to be almost negligible.

Now let us consider the variance of the estimate $\underline{\hat{H}}(X)$. The estimate $\underline{\hat{H}}$ can be expressed as a linear function of $m_i^{(r)}$ (using 5.3.11)):

$$\underline{\hat{H}} = \frac{1}{N} \sum_{r=1}^{R} \sum_{i=1}^{K} \theta_{ri} m_i^{(r)} \tag{5.4.3}$$

where $m_i^{(r)}$ obey the multinominal distribution.

Then the variance of $\hat{\underline{H}}$ can be calculated as follows:

$$V\{\hat{\underline{H}}\} = \sum_{r=1}^{R}\sum_{i=1}^{K}\theta_{ri}^2\frac{\rho_r q_i^{(r)}(1-\rho_r q_i^{(r)})}{N} - \sum_{r'=1}^{R}\sum_{r=1}^{R}\sum_{i'=1}^{K}\sum_{i=1}^{K}\theta_{ri}\theta_{r'i'}\frac{\rho_r\rho_{r'}q_i^{(r)}q_{i'}^{(r')}}{N}(1-\delta_{ri,\,r'i'}), \tag{5.4.4}$$

where

$$\delta_{ri,r'i'} = \begin{cases} 1 & \text{if } r' = r,\ i' = i \\ 0 & \text{otherwise} \end{cases}. \tag{5.4.5}$$

For the evaluation of the variance from the experimental data we should substitute the frequencies $m_i^{(r)}/N$ instead of $\rho_r q_i^{(r)}$. We obtain:

$$\hat{V}\{\hat{\underline{H}}\} = \frac{1}{N^2}\sum_{r=1}^{R}\sum_{i=1}^{K}\theta_{ri}^2 m_i^{(r)} - \frac{\hat{H}^2}{N}. \tag{5.4.6}$$

### 5.5. *Statistical verification of the subdivisions for the upper and lower bounds*

As was stated in Section 3 the improvement of the upper bound can be achieved by subdivision of the totality of linguistic situations into several groups. Some of the original groups are, in practice, very small and should be joined to larger groups with similar distributions. A modified two-sample Wilcoxon rank test with ties (using mid-ranks) (Lehmann [41]) was found to be suitable in order to verify the difference and the similarity of the probability distributions $q_i^{(r)}$ of the groups.

The problem of subdivision for the lower bound is formulated in a different way. Here we have to rely mostly on guesser's decisions about the bounds for probabilities $p_i = P_{X/S^L}(x_{k_i(j)}/s_j^L)$ and their sums (for instance, the guesser can indicate that $p_1 > 0.85$, or $p_1 < 0.85$, but $p_1 + p_2 > 0.85$, etc.). The validity of the characterization, given by the guesser, can be checked, however, by calculation of the average values $q_i^{(r)}$ of the corresponding probabilities in the groups (which should not be too close to the boundaries). It has been found that the majority of incorrect qualifications made by the guesser is connected with specific linguistic situations (for instance, the situation "after possible prefix") and can be properly corrected.

## 6. THE EXPERIMENT

### 6.1. *Introduction*

The goal of our experimental research was to investigate the information-theoretical properties of the Hebrew language using the improved prediction method.

A point of great interest was the comparison of the information-theoretical characteristics of Hebrew with those of European languages, since they differ so much in grammar and transciption. The theoretical treatment given in Sections 3–5 forms a basis for a new kind of experimental study and provides tools and methods which enable more accurate and more reliable results.

### 6.2. *The description of the experiment*

The experiment consists of 1000 independent trials: in each of them the guesser must guess a symbol which is following a text of 1000 letters, unknown to the guesser before the

experiment. The texts were taken from original modern Hebrew fiction (twentieth-century prose) and were not overlapping.

A random number generator was used to prepare a sequence of random numbers, which were later used for the choice of the symbol to be guessed. The range of the random numbers was taken with precautions in order to avoid influence of beginnings and endings of lines.

Starting from the randomly chosen symbol, a window was open in the backward direction uncovering a text of approximately 1000 letters ($\pm 10$ letters) which should be read prior to the guessing itself (thus, $L = 1000$ in our experiment). After reading the text the guesser was asked to relate the linguistic situation to one of the following groups:

1. There is a single continuation with a probability close to 1 ($p_1 \geqslant 0.85$).
2. There are two possible continuations, one at least twice as probable as the other, with a total probability $p_1 + p_2 \geqslant 0.85$.
3. There are two possible continuations with almost equal probabilities and with a total probability $p_1 + p_2 \geqslant 0.85$.
4. There are up to four possible continuations with a total probability $p_1 + p_2 + p_3 + p_4 \geqslant 0.85$.
5. All the other cases.

The guesser is also asked to state if the last symbol of the visible text can be characterized as a "possible prefix" and only then a number of attempts is made to guess the unknown symbol, starting with the most probable until the guess is correct.

A number of a priori and a posteriori characteristic was recorded as follows:

1. A priori characteristics:
   (a) The previous symbol (the last symbol of the visible text).
   (b) The number of the position of the symbol to be guessed in the word.
   (c) The characterization of the previous symbol as a "possible prefix".
2. A posteriori characteristics:
   (a) The outcome of each attempt.
   (b) The correct continuation.
   (c) The number of attempts (the symbol of the "reduced text").
   (d) The characterization of the guessed letter as "a first letter-prefix", "a first letter-stem", "a letter after prefix", "a middle letter", "an end letter".

It should be noted that in Hebrew the articles and most of the prepositions and conjunctions are connected with the following word and therefore we call them "prefixes".

In the main experiment two guessers were engaged together as a collective guessing team. They were free to discuss all the questions arising in the process of guessing. Decisions were usually made on the basis of a consensus. The guessers were aided by tables of probabilties of first letters in total, first letters in words beginning with a stem, first letters in words beginning with a prefix. The tables were prepared from the material in Choueka and Yeshurun [14] and Choueka [42].

## 7. THE EVALUATION OF THE UPPER BOUND OF ENTROPY

### 7.1. *Subdivision for the upper bound*

The primary groups of data 1–5 obtained in the experiment were further subdivided into 11 secondary groups using the a priori characterizations: "after a possible prefix" ("a") and

"a beginning of a word" ("f"). Those secondary groups are listed in Table 8 and the distributions of the number of attempts are given in Table 9.

The two-sample modified Wilcoxon sum–rank test with ties (mid-ranks) (Lehmann [41]) was used to verify similarities and differences between the various secondary groups. Most of the groups were pair-wise tested. The tests proved the groups I, J, K, which constitute the original group 5 to be different, justifying their separation. It has also been proved that group A differs from B and D, and that group D differs from E and F. Groups C, B, H, F, G and E are too small to be considered separately and they should be combined with larger groups. The final decision was based on values of estimates of probabilities that two groups may differ more than they do if they belong to the same population.

It was decided to combine groups C, D, B and H into one group and groups F, G, I and

Table 8. Secondary groups of experimental results based on primary subdivision and a priori characterization

| No. | Group | Definition of the group | Size |
|---|---|---|---|
| 1 | A | Group 1 excluding "a"s and "f"s | 442 |
| 2 | B | The "a"s of group 1 | 10 |
| 3 | C | The "f"s of group 1 | 14 |
| 4 | D | Group 2 excluding "a"s and "f"s | 84 |
| 5 | E | The "a"s of group 2 | 5 |
| 6 | F | The "f"s of group 2 | 8 |
| 7 | G | The primary group 3 | 6 |
| 8 | H | The primary group 5 | 11 |
| 9 | I | Group 5 excluding "a"s and "f"s | 124 |
| 10 | J | The "a"s of group 5 | 118 |
| 11 | K | The "f"s of group 5 | 158 |

Table 9. Subdivision into secondary groups

| Group $y_i$ | $m_i$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K |
| 1 | 446 | 7 | 12 | 62 | 2 | 4 | 2 | 7 | 62 | 37 | 60 |
| 2 | 8 | 1 | 1 | 14 | 2 | 2 | 3 | 2 | 21 | 11 | 23 |
| 3 | 3 | 1 | 0 | 4 | 0 | 0 | 1 | 1 | 7 | 3 | 13 |
| 4 | 1 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 9 | 5 | 11 |
| 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 12 | 8 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 8 | 9 |
| 7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 6 | 6 |
| 8 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 7 | 8 |
| 9 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 6 | 2 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 1 | 2 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 6 | 6 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 1 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 4 |
| 15 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 4 | 1 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Total | 462 | 10 | 14 | 84 | 5 | 8 | 6 | 11 | 124 | 118 | 158 |

E into another group. Finally, we have obtained five groups for the evaluation of the upper bound. The final subdivision is shown in Tables 10 and 11.

### 7.2. *Calculations of the estimates of the upper hand of entropy*

The final results of the calculations of the upper bound estimates for both Shannon's original method and the improved method are given in Table 12.

Table 10. The arrangement of secondary groups into five groups in accordance with results of the Wilcoxon tests

| Final group | Definition | Size ($N_r$) |
|---|---|---|
| $a_1$ | A—group 1 excluding "a"s and "f"s | 462 |
| $a_2$ | C, D, B, H—"f"s of group 1; group 2 excluding "a"s and "f"s; "a"s of 1; group 4 | 119 |
| $a_3$ | F, G, I, E—"f"s of 2; group 3; group 5 excluding "a"s and "f"s; "a"s of 2 | 143 |
| $a_4$ | K—"f"s of group 5 | 158 |
| $a_5$ | J—"a"s of group 5 | 118 |

Table 11. The final subdivision for the evaluation of the upper bound of entropy per symbol

| $y_i$ | | $m_i$ | | | | | |
|---|---|---|---|---|---|---|---|
| | Group | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | Total |
| 1 | | 446 | 88 | 70 | 60 | 37 | 701 |
| 2 | | 8 | 18 | 28 | 23 | 11 | 88 |
| 3 | | 3 | 6 | 8 | 13 | 3 | 33 |
| 4 | | 1 | 2 | 11 | 11 | 5 | 30 |
| 5 | | 1 | 0 | 4 | 8 | 12 | 25 |
| 6 | | 0 | 0 | 5 | 9 | 8 | 22 |
| 7 | | 2 | 0 | 2 | 6 | 6 | 16 |
| 8 | | 1 | 1 | 1 | 8 | 7 | 18 |
| 9 | | 0 | 2 | 1 | 2 | 6 | 11 |
| 10 | | 0 | 0 | 4 | 2 | 1 | 7 |
| 11 | | 0 | 0 | 2 | 6 | 6 | 14 |
| 12 | | 0 | 0 | 2 | 0 | 0 | 2 |
| 13 | | 0 | 1 | 1 | 1 | 3 | 6 |
| 14 | | 0 | 0 | 1 | 4 | 2 | 7 |
| 15 | | 0 | 1 | 2 | 1 | 4 | 8 |
| 16 | | 0 | 0 | 0 | 2 | 2 | 4 |
| 17 | | 0 | 0 | 0 | 0 | 2 | 2 |
| 18 | | 0 | 0 | 1 | 1 | 1 | 3 |
| 19 | | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | | 0 | 0 | 0 | 1 | 1 | 2 |
| 21 | | 0 | 0 | 0 | 0 | 1 | 1 |
| Total | | 462 | 119 | 143 | 158 | 118 | 1000 |

Table 12. Final results for the estimate of the upper bound of entropy in nats and bits

| | | The estimate | The bias (abs. value) | The s.d. | Final results |
|---|---|---|---|---|---|
| Shannon's estimate | Nats | 1.327 | 0.008 | 0.050 | $1.335 \pm 0.050$ |
| $\hat{H}_{\mathrm{Sh}}(X)$ | Bits | 1.915 | 0.012 | 0.072 | $1.927 \pm 0.072$ |
| | Nats | 1.078 | 0.044 | 0.044 | $1.122 \pm 0.044$ |
| $\hat{\hat{H}}(X)$ | Bits | 1.555 | 0.063 | 0.063 | $1.618 \pm 0.063$ |

## 8. THE EVALUATION OF THE LOWER BOUND OF ENTROPY

### 8.1. *Subdivision for the lower bound; determination of boundaries of convex domains*

In order to apply experimentally the technique which was developed in Sections 4 and 5 we should manage to subdivide the population of linguistic situations into several groups such that the conditional probability vectors $\mathbf{p}_j$ for all the cases belonging to a certain group are concentrated in a corresponding convex domain which is a subset of the simplex $\mathbb{C}$ of all the ordered probability vectors. Thus, such a subdivision of the linguistic situations induces a subdivision of the simplex $\mathbb{C}$ into convex subsets (which may, in general, overlap) by hyperplanes, defined by equations of the following general form:

$$\sum_{i=1}^{K} c_i p_i = \text{const.} \tag{8.1.1}$$

These hyperplanes together with the boundaries of the simplex $\mathbb{C}$ itself form the boundaries of the convex domains corresponding to different groups of linguistic situations.

Note that the characterizations of the linguistic cases which were given by the guesser according to criteria 1–5 in Section 6.1 separate the different groups of cases just by conditions of the form (8.1.1). Hence, the hyperplanes subdividing the simplex $\mathbb{C}$ may be specified by equations:

$$p_1 = p_0, \tag{8.1.2}$$

$$p_1 + p_2 = p_0, \tag{8.1.3}$$

$$p_1 = 2p_2 \quad \left(\text{or} \quad \frac{p_1}{p_2} = 2\right), \tag{8.1.4}$$

$$p_1 + p_2 + p_3 + p_4 = p_0. \tag{8.1.5}$$

(In our experiment $p_0 = 0.85$.)

It is natural to expect that the characterization of the linguistic cases with respect to the boundaries (8.1.2)–(8.1.5) is subjected to errors: some of the cases have probability vectors which lie, in fact, outside the domain indicated by the guesser. The values of the sums of frequencies [or, for (8.1.4), the ratio of frequencies] for different groups of linguistic situations (which are estimates for the sums of average conditional probabilities for the group) can serve as "indicators" of the validity of the characterization for the group as a whole: they should take values which are located well inside the corresponding domain and not close to the boundaries separating this domain from its neighbours. An analysis of the secondary groups formed by use of the a priori characteristics (Table 9) shows that some small groups (such as B, E, F and G) do not satisfy the criteria which they should obey according to the original characterization of the guesser: the sums of the corresponding frequencies take values lying outside the proper domains. This means that a large part of erroneously characterized cases belong to those groups. Of course, the values of the frequencies are subjected to statistical fluctuations. In any case in order to obtain a more reliable result we should for such doubtful cases always make a decision which will provide a *lower* value of the lower bound. It can be seen that because of the concavity of the entropy function and the convexity of the lower bound (Theorem 4.3.1) it is more reliable to join such a small group to a group whose domain has vertices with larger values of entropy.

Following these ideas we finally come to four convex domains $C_1$, $C_2$, $C_3$ and $C_4$ in the

simplex © and to corresponding groups of linguistic situations ($L$-grams) $d_1$, $d_2$, $d_3$ and $d_4$, as follows:

$C_1$—the domain defined by the expressions:

$$\left.\begin{array}{l} p_1 \geqslant p_2 \geqslant p_3 \geqslant \ldots \geqslant P_K \geqslant 0, \quad \sum_{i=1}^{K} p_i = 1 \\ p_1 \geqslant p_0 \end{array}\right\} \tag{8.1.6}$$

$C_2$—the domain defined by the expressions:

$$\left.\begin{array}{l} p_1 \geqslant p_2 \geqslant \ldots \geqslant p \geqslant 0, \quad \sum_{i=1}^{K} p_1 = 1 \\ p_1 \leqslant p_0, \quad p_1 + p_2 \geqslant p_0, \quad p_1 \geqslant 2p_2 \end{array}\right\} \tag{8.1.7}$$

$C_3$—the domain defined by the expressions:

$$\left.\begin{array}{l} p_1 \geqslant p_2 \geqslant \ldots \geqslant p_K \geqslant 0, \quad \sum_{i=1}^{K} p_i = 1 \\ p_1 \leqslant p_0, \quad p_1 + p_2 + p_3 + p_4 \geqslant p_0 \end{array}\right\} \tag{8.1.8}$$

$C_4$—the domain defined by the expressions:

$$\left.\begin{array}{l} p_1 \geqslant p_2 \geqslant \ldots \geqslant p_K \geqslant 0 \ , \quad \sum_{i=1}^{k} p_i = 1 \\ p_1 + p_2 \leqslant p_0 \end{array}\right\}. \tag{8.1.9}$$

The correspondence between the secondary groups of $L$-grams, the final groups and the convex domains, as well as the values of indicators, are given in Table 13. Table 14 shows the distribution of cases in the final groups. The smoothed distributions obtained according to Section 5.3 are given in Table 15.

### 8.2. *Final results for the estimate of the lower bound of entropy*

Applying the methods and techniques developed in Sections 4 and 5, we come to the following final results for the lower bound of entropy of printed Hebrew:

$$\underline{\hat{H}}(X) = 1.217 \pm 0.059 \text{ bits}$$

$$\underline{\hat{H}}_{\text{Sh}}(X) = 1.019 \pm 0.053 \text{ bits}.$$

## 9. RESULTS AND DISCUSSIONS

### 9.1. *Entropy and redundancy of printed Hebrew and comparison with other results*

The experiment described in Sections 6–8 gives estimates of the upper and lower bounds of the "entropy of 1000th order" $F_{1000}$ which certainly can be taken as estimates of the upper and lower bounds of the actual entropy per symbol $H$. Finally, we have:

$$\hat{\bar{H}} = 1.618 \pm 0.063 \text{ bits} \tag{9.1.1}$$

$$\underline{\hat{H}} = 1.217 \pm 0.059 \text{ bits}. \tag{9.1.2}$$

Table 13. Final groups, final convex domains and total values of the relevant indicators for the evaluation of the lower bound

| Final group | Final convex domain | Total values of indicators | | | | Secondary groups |
|---|---|---|---|---|---|---|
| | | $\hat{q}_1$ | $\hat{q}_1 + \hat{q}_2$ | $\hat{q}_1 + \hat{q}_2 + \hat{q}_3 + \hat{q}_4$ | $\hat{q}_1/\hat{q}_2$ | |
| $d_1$ | $C_1$ | 0.962 | | | | A + C |
| $d_2$ | $C_2$ | 0.738 | 0.905 | | 4.43 | D |
| $d_3$ | $C_3$ | 0.55 | | 0.925 | | B + E + F + G + H |
| $d_4$ | $C_4$ | | 0.535 | | | I + J + K |

Table 14. The final subdivision for the evaluation of the lower bound

| $y_i$ Group | $m_i$ | | | | $m_i$ (total) |
|---|---|---|---|---|---|
| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | |
| 1 | 458 | 62 | 22 | 159 | 701 |
| 2 | 9 | 14 | 10 | 55 | 88 |
| 3 | 3 | 4 | 3 | 23 | 33 |
| 4 | 1 | 2 | 2 | 25 | 30 |
| 5 | 1 | 0 | 1 | 23 | 25 |
| 6 | 0 | 0 | 0 | 22 | 22 |
| 7 | 2 | 0 | 0 | 14 | 16 |
| 8 | 1 | 1 | 0 | 16 | 18 |
| 9 | 1 | 0 | 1 | 9 | 11 |
| 10 | 0 | 0 | 0 | 7 | 7 |
| 11 | 0 | 0 | 0 | 14 | 14 |
| 12 | 0 | 0 | 0 | 2 | 2 |
| 13 | 0 | 0 | 1 | 5 | 6 |
| 14 | 0 | 0 | 0 | 7 | 7 |
| 15 | 0 | 1 | 0 | 7 | 8 |
| 16 | 0 | 0 | 0 | 4 | 4 |
| 17 | 0 | 0 | 0 | 2 | 2 |
| 18 | 0 | 0 | 0 | 3 | 3 |
| 19 | 0 | 0 | 0 | 0 | 0 |
| 20 | 0 | 0 | 0 | 2 | 2 |
| 21 | 0 | 0 | 0 | 1 | 1 |
| Total | 476 | 84 | 40 | 400 | 1000 |

Table 15. The smoothed distributions in the final groups for the evaluation of the lower bound

| Group $y_i$ | $\check{m}_i$ | | | | $\check{m}_i$ (total) |
|---|---|---|---|---|---|
| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | |
| 1 | 458 | 62 | 22 | 159 | 701 |
| 2 | 9 | 14 | 10 | 55 | 88 |
| 3 | 3 | 4 | 3 | 24 | 33 |
| 4 | 1 | 2 | 2 | 24 | 30 |
| 5 | 1 | 1/4 | 1 | 23 | 25 |
| 6 | 1 | 1/4 | 1/4 | 22 | 22 |
| 7 | 1 | 1/4 | 1/4 | 15 | 17 |
| 8 | 1 | 1/4 | 1/4 | 15 | 17 |
| 9 | 1 | 1/7 | 1/4 | 10 | 11 |
| 10 | 0 | 1/7 | 1/4 | 10 | 10.5 |
| 11 | 0 | 1/7 | 1/4 | 10 | 10.5 |
| 12 | 0 | 1/7 | 1/4 | 21/4 | 5.75 |
| 13 | 0 | 1/7 | 1/4 | 21/4 | 5.75 |
| 14 | 0 | 1/7 | 0 | 21/4 | 5.75 |
| 15 | 0 | 1/7 | 0 | 21/4 | 5.75 |
| 16 | 0 | 0 | 0 | 4 | 4 |
| 17 | 0 | 0 | 0 | 5/2 | 2.5 |
| 18 | 0 | 0 | 0 | 5/2 | 2.5 |
| 19 | 0 | 0 | 0 | 1 | 1 |
| 20 | 0 | 0 | 0 | 1 | 1 |
| 21 | 0 | 0 | 0 | 1 | 1 |
| Total | 476 | 84 | 40 | 400 | 1000 |

The results show a considerable improvement in comparison with the upper and lower bounds calculated from the same experimental data by the original Shannon method:

$$\hat{\hat{H}}_{\mathrm{Sh}} = 1.927 \pm 0.072 \text{ bits} \tag{9.1.3}$$

$$\underline{\hat{H}}_{\mathrm{Sh}} = 1.019 \pm 0.053 \text{ bits.} \tag{9.1.4}$$

Thus the gap between the upper and the lower bounds is reduced more than twice by the use of the improved method:

$$\delta\hat{H}(X) = \hat{\bar{H}}(X) - \underline{\hat{H}}(X) = 0.401 \text{ bits} \tag{9.1.5}$$

$$\delta\hat{H}_{\mathrm{Sh}}(X) = \hat{\bar{H}}_{\mathrm{Sh}}(X) - \underline{\hat{H}}_{\mathrm{Sh}}(X) = 0.908 \text{ bits} \tag{9.1.6}$$

The results (9.1.1) and (9.1.2) show, in particular, that the statistical inaccuracy is rather small (4–5%) in our experiment, so that the accuracy of the evaluation of the entropy per symbol is limited mostly by the gap between the upper and the lower bounds. A theoretical analysis of the limits of the gap is given in Reingold [8], Ch. 3.

The amount of information in the preceding text about the following symbol [cf. (2.1.7)] is bounded for Hebrew by:

$$\hat{\bar{I}}(S, X) = F_1(X) - \underline{H}(X) = 2.767 \pm 0.070 \text{ bits} \tag{9.1.7}$$

$$\underline{\hat{I}}(S, X) = F_1(X) - \bar{H}(X) = 2.366 \pm 0.074 \text{ bits.} \tag{9.1.8}$$

The bounds of the redundancy of a language are defined by formulae [cf. (2.1.8)]:

$$\hat{\bar{R}} = \frac{F_0 - \hat{H}}{F_0}; \qquad \underline{\hat{R}} = \frac{F_0 - \hat{\bar{H}}}{F_0}. \tag{9.1.9}$$

The zero-order entropy for Hebrew is $F_0 = \log_2 23 = 4.52$ bits.

Using (9.1.1) and (9.1.2), we obtain:

$$\hat{\bar{R}} = 73\%; \qquad \underline{\hat{R}} = 64\%. \tag{9.1.10}$$

It can be seen that the entropy per symbol of Hebrew is considerably larger (by $\approx 0.3$–$0.5$ bits) than that of European languages, Armenian and Kazakh. The redundancy of the European languages is larger by approximately 10% (see Table 3) than that of Hebrew.

In order to understand the origin of these differences some special features of printed Hebrew text should be considered. The printed Hebrew has a partly syllabic structure: the vowels "a" and "e" are not indicated by special letters, the vowel "o" and "u" are denoted by the same letter, and moreover, this letter and the letter for "i" are sometimes omitted in older texts, which are written in the so-called "ktav khaser" ("missing spelling") in contrast with modern texts written in "ktav maleh" ("full spelling"). In our experiment, twentieth-century Hebrew texts were used which included both types of spelling. The partial removal of vowels results, of course, in a reduction of redundancy and it can explain the higher entropy per symbol of Hebrew in comparison with other languages using complete transcription of vowels.

## 9.2. *Entropy per symbol for special linguistic cases*

It is known that information is distributed in a text in a highly non-uniform way. The uncertainty of a symbol following a given $L$-gram depends very much on the position of the symbol in the word and on other grammatical features of the situation.

Subjectively, the beginnings of words seem to constitute the most difficult situations for a guesser in such languages as English or Russian. However, it should be noted that in Hebrew the articles and most of the prepositions and conjunctions are connected with the following words. Therefore, it is worth considering separately this type of situation, which we called "a letter after a prefix" (a posteriori). Other cases of interest are "a middle letter" (excluding the cases "after prefix"), "an end letter" and "space".

Table 16 gives the upper and lower bounds of the entropy of the first letter of a word (for "beginning with a prefix", "beginning with a stem" and "total", respectively).

These results are compared with the values of the first order entropy $F_1^f$ for the same types of situations. Similar results for Russian obtained by Piotrovski [20] are also given. It can be learned from the table that "the beginning with a stem" provides considerably more information than "the beginning with a prefix".

The results for the "beginning with a stem" are closer to the "total" results for Russian, apparently because of the fact that there are no such "prefixes" as articles, prepositions and conjunctions in Russian.

Table 17 contains the estimates for the upper and lower bounds of entropy for the situations "after prefix", "middle letter", "end letter" and "space". It is seen that the entropy of letters "after prefix" is the highest one and even higher than the entropy of a "first letter-stem".

This situation is specific for Hebrew in contrast with such languages as English and Russian. The entropy of "middle letters" is a little less than the general entropy per letter. The entropy of "end letters" is, as expected, much less, and the entropy of spaces is close to zero.

### 9.3. *Applications of the method and the results*

The method developed in the present work can be easily applied to other languages (though the characterization of the groups and the a priori characteristics recorded should be adapted to each language individually). A more general field of application of the method is an information-theoretical study of other forms of information-carrying processes which possess a very complicated probabilistic structure and interdependence between their elements, inaccessible for a direct statistical analysis, but appeal to our experience, knowledge and intuition. Objects such as meaningful images, music, languages on other linguistic levels (such as syllables, morphems, words, etc.), information exchange between a human group and a leader, systems with man–machine interaction, etc., can be investigated by some modifications of the prediction method. The prediction experiment itself can be used in psychology as a test of personality features: intellectuality, memory, language knowledge, ability to evaluate probabilities of random events, decision-making properties, etc., for which purpose a set of standard texts should be prepared.

The results of the experiment can be used in communication engineering, as was indicated by Shannon [2], and in cryptography (for the preliminary elimination of source

Table 16. Entropy of the first letter of a word in bits

| | First letter-prefix | First letter-stem | First letter-total | First letter-total for Russian |
|---|---|---|---|---|
| $\hat{\bar{H}}$ | 2.47 | 3.27 | 2.94 | 3.45 |
| $\underline{\hat{H}}$ | 1.58 | 2.39 | 1.98 | 2.98 |
| $F_1^{(f)}$ | 2.93 | 4.24 | 3.88 | 4.23 |

Table 17. Entropy per symbol for different types of linguistic situations

| | After prefix | Middle letter | End letter | Space |
|---|---|---|---|---|
| $\hat{\bar{H}}$ | 3.57 | 1.48 | 1.01 | 0.176 |
| $\underline{\hat{H}}$ | 2.73 | 0.87 | 0.471 | 0.0816 |

redundancy). Some applications of such results in linguistics were discussed by Bar-Hillel [43], Herdan [5], Piotrovski [20, 22], and others (Yaglom *et al.* [4]; Dobrushin [44]; Kondratov [7]).

*Acknowledgements*—The authors feel very much indebted to Michael Dror (Tel-Aviv) for his crucial contributions by participation in the prediction experiments. They are thankful to Y. Choueka (Bar-Ilan University, Ramat-Gan) and S. P. Ladany (Ben-Gurion University, Beer-Sheba) for the valuable material on statistics of Hebrew language, to V. Raskin (Purdue University, West Lafayette, USA) and H. Konijn (Tel-Aviv University, Ramat-Aviv) for interesting discussions. They appreciate the work done by the editor, W. Ebeling (Humboldt University, Berlin), and by the referee who helped to improve and update the paper.

## REFERENCES

1. C. E. Shannon, A mathematical theory of communications, *Bell Syst. tech. J.* **27**, 379–423 (1948).
2. C. E. Shannon, Prediction and entropy of printed English, *Bell Syst. tech. J.* **30**, 50–64 (1951).
3. B. Mandelbrot, An informational theory of the statistical structure of a language, in *Communication Theory*, edited by W. Jackson, Academic Press, New York, pp. 486–502 (1953).
4. A. M. Yaglom, I. M. Yaglom and R. L. Dobrushin, Information theory and linguistics, *Voprosy Yazyoznaniya* (*Problems of Linguistics*), **1**, 100–110 (1960).
5. G. Herdan, *The Advanced Theory of Language as Choice and Chance*. Springer, Berlin (1966).
6. N. Rychkova, Linguistics and mathematics, *Nauka Zhizn* (*Science and Life*), No. 9, 76–77 (1961).
7. A. M. Kondratov, Information theory and poetics (The entropy of the rhythm of the Russian language), *Probl. Cybern.* **9**, 279–286 (1963).
8. Z. Reingold, Evaluation of the entropy of a language by an improved prediction method with application to printed Hebrew, M. Sc. thesis, Tel-Aviv University (1980).
9. W. Hillberg, Der bekannte Grenzwert der redundanzfreien Information in Texten—eine Fehlinterpretation der Shanonschen Experimente? (The well-known lower bound of information in written language—is it a misinterpretation of Shannon's experiments?) *Frequenz* **44**, 243–248 (1990).
10. W. Ebeling and G. Nicolis, World frequency and entropy of symbolic sequences: a dynamic perspective, *Chaos, Solitons & Fractals* **2**, 635–650 (1992).
11. J. S. Nicolis, A. A. Katsikas, Chaotic dynamics of linguistic-like processes on the syntactical and semantic levels: in the pursuit of a multifractal generator, in *Studies in Nonlinearity in Life Sciences*, ed. by B. West, World Scientific, Singapore (1992).
12. K. Kuepfmueller, The entropy of the German language, *Fernmeldetechnische Zeitschrift, (J. Telecommun.)* **VII**, 265–272 (1954).
13. S. P. Ladany, Valid data for design of Hebrew language information processing equipment, in *Proceedings of the 37th ASIS Annual Meeting*, **11**, Washington DC (1974).
14. Y. Choueka and S. Yeshurun, Statistical aspects of modern Hebrew prose, in *Proceedings of the 5th National Conference of IPA* (*Information Processing Association of Israel*), Jerusalem (1969).
15. P. Grassberg, Estimating the information context of symbol sequences and efficient codes. *IEEE Trans. Inform. Theory* **IT-35**, 669–675 (1989).
16. N. G. Burton and J. C. R. Licklider, Long range constraints in the statistical structure of printed English, *Amer. J. Psychol.* **68**, 650–653 (1955).
17. A. A. Piotrovskaia, R. G. Piotrovski and K. A. Razzhivin, The entropy of the Russian language, *Voprosy Yazykoznaniya* (*Problems of Linguistics*) No. 6, 115–130 (1962).
18. A. M. Yaglom and J. M. Yaglom, *Probability and Information* (3rd revised Edn). Science Publishing House, Leningrad (1973).
19. T. M. Cover and R. C. King, A convergent gambling estimate of the entropy of English, *IEEE Trans. Inform. Theory* **IT-24**, 413–421 (1978).
20. R. G. Piotrovski, *Informational Measurements of Language*. Science Publishing House, Leningrad (1968).
21. R. G. Piotrovski, Entropy and redundancy of four European languages, in *Statistical Methods in Linguistics* Vol. 5, Stockholm (1969).
22. R. G. Piotrovski, *Text, Computer, Man*. Science Publishing House, Leningrad (1975).
23. N. Petrova, R. G. Piotrovski and R. Giraud, The entropy of written French, *Bull. Soc. Linguist. Paris* **59**, 130–152 (1964).
24. G. P. Boguslavskaia and L. A. Novak, The entropy of the English and Rumanian languages, *Statistics of Speech* II, Minsk (1968).
25. G. Boguslavskaia, T. Koženec and R. G. Piotrovski, Informational estimates of text, *ZPhSK* (*J. Phonetics, Linguistics Comm. Res.*) **24**, (1970).
26. L. A. Novak and R. G. Piotrovski, A prediction experiment for the entropy of Rumanian language, in *Statistical Linguistics*, edited by C. Tagliavani, Vol. 3. Bologna (1971).
27. I. A. Korolenko, I. V. Matkovski, L. A. Novak and R. G. Piotrovski, The entropy of Rumanian and Moldavian texts, in *Coordinative Conference on Comparative and Typological Studies of Rumanian Languages*, Leningrad (1964).

28. D. A. Baytanaieva and K. B. Bektaiev, The entropy of Kazakh text, SKT (Statistics of Kazakh Text), Issue III, Alma-Ata (1973).
29. R. A. Kazarian, The evaluation of the entropy of the Armenian text, *News Acad. Sci. Armenia* (*The Physical and Mathematical Sciences*) **14**, 161–173 (1961).
30. D. N. Lenskoi, On the evaluation of the entropy of Adyghe printed texts, in *Scientific Notes of the Kabardino–Balkarskii University* (*The Phys. and Math. Series*) Nalchik, Issue 16, pp. 165–166 (1962).
31. L. Doležel, Prediction of the entropy and redundancy of written Czech, *Slovo i Slovestnost* **24**, 165–175 (1963).
32. A. P. Savchuk, Experimental evaluation of the entropy of Russian language, *Conference of Applications of Mathematical Methods to the Study of Language in Fiction*. USSR Academy of Sciences, Gorkii (1961).
33. A. V. Gut, The entropy of the printed Polish text, *M.Sc. thesis*, Leningrad State University, Leningrad (1966).
34. Ch. C. Georgiev, Information measurements of the Bulgarian language. Ph.D. thesis, Leningrad (1973).
35. R. M. Fano, *Transmission of Information*, MIT Press, Cambridge, and Wiley, New York (1960).
36. N. Bourbaki, *Topologie Generale, Elements de Mathematique*, Partie I, Livre III. Hermann, Paris (1960).
37. G. B. Danzig, *Linear Programming and Extensions*. Princeton Unviersity Press, Princeton, New Jersey (1963).
38. G. P. Basharin, On a statistical estimate for the entropy of a sequence of independent random variables, *Theory Prob. Appl.* **4**, 333–336 (1959).
39. L. B. Levitin and Z. Reingold, An improved estimate for the entropy of a discrete random variable, The Annual Meeting of the Israel Statistical Association, Tel-Aviv University, Tel-Aviv (1978).
40. B. L. Van der Waerden, *Mathematische Statistik*. Springer, Berling (1957).
41. E. L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*. Holden–Day, San Francisco (1975).
42. Y. Choueka, Private communication (1979).
43. Y. Bar–Hillel, *Language and Information* (selected essays on their theory and applications). Addison-Wesley and Academic Press, Jerusalem (1964).
44. R. L. Dobrushin, Mathematical methods in linguistics, *Math. Education* (new series), No. 6, pp. 37–60. Phizmatgiz, Moscow (1961).
45. V. A. Garmash, N. E. Kirillov and D. S. Lebedev, An experimental study of statistical properties of message sources, *Problemy Pered. Inf.* (*Book: Problems of Information Transmission*), Issue 5, 000–000, Moscow (1960).
46. G. A. Barnard, Statistical calculation of word entropies for four western languages, *IRE Trans. Inform. Theory* **IT-1**, 49–53 (1955).
47. S. P. Ladany, The structure of the printed Hebrew language and its efficiency in transmitting information, *Hebrew Comput. Linguist* **4**, 69–82 (1971).
48. 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* **IT-22**, 123–000 (1978).
49. R. Manfrino, Printed Portugese entropy: statistical calculations, *IEEE Trans. Inform. Theory* **IT-16**, 122–000 (1970).
50. L. B. Levitin and Z. Reingold, Evaluation of the entropy of printed Hebrew by an improved prediction method, in *Proceedings of the IEEE Conference*, Tel-Aviv, Israel (1979).