Prediction and Entropy of Printed English

By C. E. SHANNON

(*ManuscriptReceived Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

1. INTRODUCTION

IN A previous paper[1] the entropy and redundancy of a language have been defined. The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each letter of a text in the language. If the language is translated into binary digits (0 or 1) in the most efficient way, the entropy H is the average number of binary digits required per letter of the original language. The redundancy, on the other hand, measures the amount of constraint imposed on a text in the language due to its statistical structure, e.g., in English the high frequency of the letter E, the strong tendency of H to follow T or of V to follow Q. It was estimated that when statistical effects extending over not more than eight letters are considered the entropy is roughly 2.3 bits per letter, the redundancy about *50* per cent.

Since then a new method has been found for estimating these quantities, which is more sensitive and takes account of long range statistics, influences extending over phrases, sentences, etc. This method is based on a study of the predictability of English; how well can the next letter of a text be predicted when the preceding N letters are known. The results of some experiments in prediction will be given, and a theoretical analysis of some of the properties of ideal prediction. By combining the experimental and theoretical results it is possible to estimate upper and lower bounds for the entropy and redundancy. From this analysis it appears that, in ordinary literary English, the long range statistical effects (up to 100 letters) reduce the entropy to something of the order of one bit per letter, with a corresponding redundancy of roughly 75%. The redundancy may be still higher when structure extending over paragraphs, chapters, etc. is included. However, as the lengths involved are increased, the parameters in question become more

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal,* v. 27, pp. 379-423, 623-656, July, October, 1948.

erratic and uncertain, and they depend more critically on the type of text involved.

2. ENTROPY CALCULATION FROM THE STATISTICS OF ENGLISH

One method of calculating the entropy // is by a series of approximations 7? F_1, F_2 > ''' , which successively take more and more of the statistics of the language into account and approach II as a limit. F_N may be called the N-gram entropy; it measures the amount of information or entropy due to statistics extending over N adjacent letters of text. F_N is given by[1]

$$FN = -\sum p(b_i, j) \log_2 p_{b_i}(j)$$

$$= -\sum_{i,j} p(b_i, j) \log_2 p(b_i, j) + \underset{i}{Z}\, p(b_i) \log p(b_i) \qquad (1)$$

in which: bi is a block of N-1 letters [(N-1)-gram]

j is an arbitrary letter following b_i

$p(b_i, j)$ is the probability of the N-gram b_i, j

$p_{b_i}(j)$ is the conditional probability of letter j after the block b_i, and is given by $p(b_i, j)/p(b_i)$.

The equation (1) can be interpreted as measuring the average uncertainty (conditional entropy) of the next letter /' when the preceding N-1 letters are known. As N is increased, F_N includes longer and longer range statistics and the entropy, II, is given by the limiting value of F_N as $N \to \infty$:

$$II = \lim_{N\to\infty} F_N . \qquad (2)$$

The N-gram entropies F_N for small values of N can be calculated from standard tables of letter, digram and trigram frequencies.[2] If spaces and punctuation are ignored we have a twenty-six letter alphabet and F_0 may be taken (by definition) to be $\log_2 26$, or 4.7 bits per letter. F_1 involves letter frequencies and is given by

$$F_1 = -\underset{i=1}{\overset{26}{X)}}\, p(i) \log_2 p(i) = 4.14 \text{ bits per letter.} \qquad (3)$$

The digram approximation F_2 gives the result

$$F_2 = -\sum_{i,j} p(i,j) \log_2 p_i(j)$$

$$= -\sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_i p(i) \log_2 p(i) \qquad (4)$$

$$= 7.70 - 4.14 = 3.56 \text{ bits per letter.}$$

[2] Fletcher Pratt, "Secret and Urgent," Blue Ribbon Books, 1942.

The trigram entropy is given by

$$
\begin{aligned}
F_3 &= -\sum_{i,j,k} p(i,j,k) \log, p_{ij}(k) \\
&= -\mathop{\mathrm{X}}_{i,j,k} p(i,j,k) \log_2 p(i,j,k) + \sum_{i,j} p(i,j) \log, p(i,j) \qquad (5) \\
&= 11.0 - 7{,}7 = 3.3
\end{aligned}
$$

In this calculation the trigram table[2] used did not take into account trigrams bridging two words, such as WOW and OWO in TWO WORDS. To compensate partially for this omission, corrected trigram probabilities $p(i, j, k)$ were obtained from the probabilities $p'(i, j, k)$ of the table by the following rough formula:

$$p(i, j, k) = \frac{2.5}{4.5} p'(i, j, k) + \frac{1}{4.5} r(i)p(jk) + \frac{1}{4.5} p(i, j)s(k)$$

where $r(i)$ is the probability of letter i as the terminal letter of a word and $s(k)$ is the probability of k as an initial letter. Thus the trigrams within words (an average of 2.5 per word) are counted according to the table; the bridging trigrams (one of each type per word) are counted approximately by assuming independence of the terminal letter of one word and the initial digram in the next or vice versa. Because of the approximations involved here, and also because of the fact that the sampling error in identifying probability with sample frequency is more serious, the value of F_3 is less reliable than the previous numbers.

Since tables of .Y-gram frequencies were not available for $N > 3$, F_4, F_5, etc. could not be calculated in the same way. However, word frequencies have been tabulated[3] and can be used to obtain a further approximation. Figure 1 is a plot on log-log paper of the probabilities of words against frequency rank. The most frequent English word "the" has a probability .071 and this is plotted against 1. The next most frequent word "of" has a probability of .034 and is plotted against 2, etc. Using logarithmic scales both for probability and rank, the curve is approximately a straight line with slope — 1 ; thus, if p_n is the probability of the rath most frequent word, we have, roughly

$$p_n = \frac{.1}{n}. \qquad (6)$$

Zipf[4] has pointed out that this type of formula, $p_n = k/n$, gives a rather good approximation to the word probabilities in many different languages. The

[3] G. Dewey, "Relative Frequency of English Speech Sounds," Harvard University Press, 1923.

[4] G. K. Zipf, "Human Behavior and the Principle of Least Effort," Addison-Wesley Press, 1949.

formula (6) clearly cannot hold indefinitely since the total probability Σp_n must be unity, while $\sum_{1}^{\infty} .1/n$ is infinite. If we assume (in the absence of any b-tter estimate) that the formula $p_n = .1/n$ holds out to the n at which the

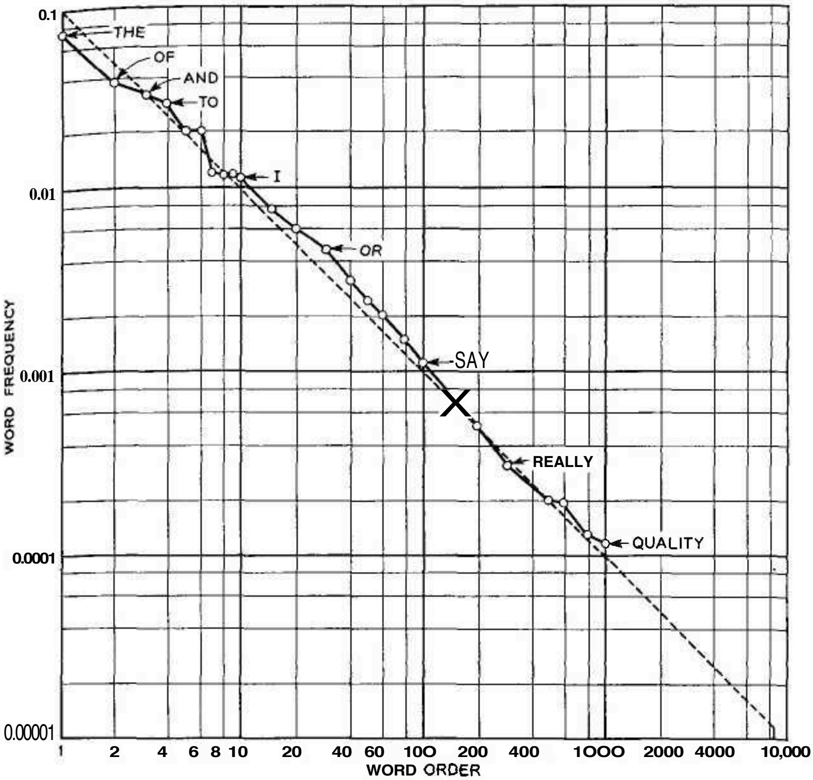

Fig. 1—Relative frequency against rank for English words.

total probability is unity, and that p_n — 0 for larger n, we find that the critical n is the word of rank 8,727. The entropy is then:

$$-\sum_{1}^{8727} pn \log_2 p_n = 11.82 \text{ bits per word,} \qquad (7)$$

or 11,82/4.5 = 2.62 bits per letter since the average word length in English is 4.5 letters. One might be tempted to identify this value with $F_{4.5}$, but actually the ordinate of the F_N curve at N = 4.5 will be above this value. The reason is that F_4 or F_5 involves groups of four or five letters regardless of word division. A word is a cohesive group of letters with strong internal statistical influences, and consequently the N-grams within words are restricted than those which bridge words. The effect of this is that we have obtained, in 2.62 bits per letter, an estimate which corresponds more nearly to, say, F_5 or F_6.

A similar set of calculations was carried out including the space as an additional letter, giving a 27 letter alphabet. The results of both 26- and 27-letter calculations are summarized below:

	F_0	F_1	F_2	F_3	F_{word}
26 letter.	4.70	4.14	3.56	3.3	2.62
27 letter.	4.76	4.03	3.32	3.1	2.14

The estimate of 2.3 for F_8, alluded to above, was found by several methods, one of which is the extrapolation of the 26-letter series above out to that point. Since the space symbol is almost completely redundant when sequences of one or more words are involved, the values of F_N in the 27-letter case will be $\frac{4.5}{5.5}$ or .818 of F_N for the 26-letter alphabet when N is reasonably large.

3. PREDICTION OF ENGLISH

The new method of estimating entropy exploits the fact that anyone speaking a language possesses, implicitly, an enormous knowledge of the statistics of the language. Familiarity with the words, idioms, cliches and grammar enables him to fill in missing or incorrect letters in proof-reading, or to complete an unfinished phrase in conversation. An experimental demonstration of the extent to which English is predictable can be given as follows: Select a short passage unfamiliar to the person who is to do the predicting. He is then asked to guess the first letter in the passage. If the guess is correct he is so informed, and proceeds to guess the second letter. If not, he is told the correct first letter and proceeds to his next guess. This is continued through the text. As the experiment progresses, the subject writes down th correct text up to the current point for use in predicting future letters. Th result of a typical experiment of this type is given below. Spaces were included as an additional letter, making a 27 letter alphabet. The first line is the original text; the second line contains a dash for each letter correctly guessed. In the case of incorrect guesses the correct letter is copied in th second line.

```
(1) THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG          (8)
(2)————ROO—————NOT-V————I. . . . . .SM———OBL————
(1) READING LAMP ON THE DESK SHED GLOW ON
(2) REA . . . . . . . . . .O—————D————SHED-GLO--O--
(1) POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
(2) P-L-S—————O---BU--L-S--O . . . . .SH . . . . RE--C . . . . . .
```

Of a total of 129 letters, 89 or 69% were guessed correctly. The errors, as would be expected, occur most frequently at the beginning of words and syllables where the line of thought has more possibility of branching out. It might be thought that the second line in (8), which we will call the *reduced text*, contains much less information than the first. Actually, both lines contain the same information in the sense that it is possible, at least in principle, to recover the first line from the second. To accomplish this we need an identical twin of the individual who produced the sequence. The twin (who must be mathematically, not just biologically identical) will respond in the same way when faced with the same problem. Suppose, now, we have only the reduced text of (8). We ask the twin to guess the passage. At each point we will know whether his guess is correct, since he is guessing the same as the first twin and the presence of a dash in the reduced text corresponds to a correct guess. The letters he guesses wrong are also available, so that at each stage he can be supplied with precisely the same information the first twin had available.

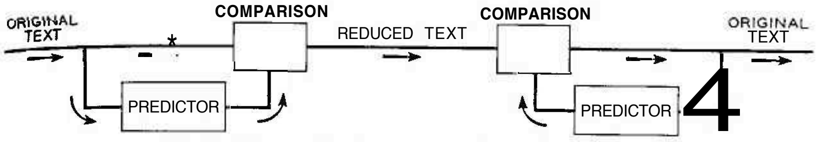

Fig. 2—Communication system using reduced text.

The need for an identical twin in this conceptual experiment can be eliminated as follows. In general, good prediction does not require knowledge of more than N preceding letters of text, with N fairly small. There are only a finite number of possible sequences of N letters. We could ask the subject to guess the next letter for each of these possible N-grams. The complete list of these predictions could then be used both for obtaining the reduced text from the original and for the inverse reconstruction process.

To put this another way, the reduced text can be considered to be an encoded form of the original, the result of passing the original text through a reversible transducer. In fact, a communication system could be constructed in which only the reduced text is transmitted from one point to the other. This could be set up as shown in Fig. 2, with two identical prediction devices.

An extension of the above experiment yields further information concerning the predictability of English. As before, the subject knows the text up to the current point and is asked to guess the next letter. If he is wrong, he is told so and asked to guess again. This is continued until he finds the correct letter. A typical result with this experiment is shown below. The first line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

```
(1) T H E R E   I S   N O   R E V E R S E   O N   A   M O T O R C Y C L E   A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
(1) F R I E N D   O F   M I N E   F O U N D   T H I S   O U T
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
(1) R A T H E R   D R A M A T I C A L L Y   T H E   O T H E R   D A Y
(2) 4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1        (9)
```

Out of 102 symbols the subject guessed right on the first guess 79 times, on the second guess 8 times, on the third guess 3 times, the fourth and fifth guesses 2 each and only eight times required more than five guesses. Results of this order are typical of prediction by a good subject with ordinary literary English. Newspaper writing, scientific work and poetry generally lead to somewhat poorer scores.

The reduced text in this case also contains the same information as the original. Again utilizing the identical twin we ask him at each stage to guess as many times as the number given in the reduced text and recover in this way the original. To eliminate the human element here we must ask our subject, for each possible N-gram of text, to guess the most probable next letter, the second most probable next letter, etc. This set of data can then serve both for prediction and recovery.

Just as before, the reduced text can be considered an encoded version of the original. The original language, with an alphabet of 27 symbols, A, B, • • • , Z, space, has been translated into a new language with the alphabet 1, 2, • • • , 27. The translating has been such that the symbol 1 now has an extremely high frequency. The symbols 2, 3, 4 have successively smaller frequencies and the final symbols 20, 21, • • • , 27 occur very rarely. Thus the translating has simplified to a considerable extent the nature of the statistical structure involved. The redundancy which originally appeared in complicated constraints among groups of letters, has, by the translating process, been made explicit to a large extent in the very unequal probabilities of the new symbols. It is this, as will appear later, which enables one to estimate the entropy from these experiments.

In order to determine how predictability depends on the number N of preceding letters known to the subject, a more involved experiment was carried out. One hundred samples of English text were selected at random from a book, each fifteen letters in length. The subject was required to guess the text, letter by letter, for each sample as in the preceding experiment. Thus one hundred samples were obtained in which the subject had available 0, 1, 2, 3, • • • , 14 preceding letters. To aid in prediction the subject made such use as he wished of various statistical tables, letter, digram and trigram tables, a table of the frequencies of initial letters in words, a list of the frequencies of common words and a dictionary. The samples in this experiment were from "*Jefferson the Virginian*" by Dumas Malone. These results, together with a similar test in which 100 letters were known to the subject, are summarized in Table I. The column corresponds to the number of preceding letters known to the subject plus one; the row is the number of the guess. The entry in column N at row S is the number of times the subject guessed the right letter at the Sth guess when (N-1)letters were known. For example,

TABLE I

	i	2	3	4	5	6	7	8	9	10	11	12	13	14	15	100
1	18.2	29.2	36	47	51	58	48	66	66	67	62	58	66	72	60	80
2	10.7	14.8	20	18	13	19	17	15	13	10	9	14	9	6	18	7
3	8.6	10.0	12	14	8	5	3	5	9	4	7	7	4	9	5	
4	6.7	8.6	7	3	4	1	4	4	4	4	5	6	4	3	5	3
5	6.5	7.1	1	1	3	4	3	6	1	6	5	2	3			4
6	5.8	5.5	4	5	2	3	2			1	4	2	3	4	1	2
7	5.6	4.5	3	3	2	2	8		1	1	1	4	1		4	1
8	5.2	3.6	2	2	1	1	2	1	1	1	1		2	1	3	
9	5.0	3.0	4		5	1	4		2	1	1	2		1		1
10	4.3	2.6	2	1	3		3	1					2			
11	3.1	2.2	2	2	2	1			1	3		1	1	2	1	
12	2.8	1.9	4		2	1	1	1			2	1	1		1	1
13	2.4	1.5	1	1	1	1	1	1	1	1		1	1			
14	2.3	1.2		1			1					1				1
15	2.1	1.0	1	1							1	1	1			
16	2.0	.9					1			1					1	
17	1.6	.7	1		2	1	1				1		2	2		
18	1.6	.5													1	
19	1.6	.4			1	1			1		1					
20	1.3	.3		1		1	1									
21	1.2	.2														
22	.8	.1														
23	.3	.1														
24	.1	.0														
25	.1															
26	.1															
27	.1															

the entry 19 in column 6, row 2, means that with five letters known the correct letter was obtained on the second guess nineteen times out of the hundred. The first two columns of this table were not obtained by the experimental procedure outlined above but were calculated directly from the known letter and digram frequencies. Thus with no known letters the most probable symbol is the space (probability .182); the next guess, if this is wrong, should be E (probability .107), etc. These probabilities are the frequencies with which the right guess would occur at the first, second, etc., trials with best prediction. Similarly, a simple calculation from the digram table gives the entries in column 1 when the subject uses the table to best advantage. Since the frequency tables are determined from long samples of English, these two columns are subject to less sampling error than the others.

It will be seen that the prediction gradually improves, apart from son, statistical fluctuation, with increasing knowledge of the past as indicated by the larger numbers of correct first guesses and the smaller numbers of high rank guesses.

One experiment was carried out with "reverse" prediction, in which the subject guessed the letter preceding those already known. Although the task is subjectively much more difficult, the scores were only slightly poorer. Thus, with two 101 letter samples from the same source, the subject obtained the following results:

No. of guess	1	2	3	4	5	6	7	8	>8
Forward	70	10	7	2	2	3	3	0	4
Reverse	66	7	4	4	6	2	1	2	9

Incidentally, the N-gram entropy F_N for a reversed language is equal to that for the forward language as may be seen from the second form in equation (1). Both terms have the same value in the forward and reversed cases.

4. IDEAL N-GRAM PREDICTION

The data of Table I can be used to obtain upper and lower bounds to th N-gram entropies F_N • In order to do this, it is necessary first to develop some general results concerning the best possible prediction of a languagt when the preceding N letters are known. There will be for the language a set of conditional probabilities $p_{i_1, i_2, \cdots, i_{N-1}}(j)$. This is the probability whet the $(N\text{-}1)$gram $i_1, i_2, \bullet\bullet\bullet, i_{N-1}$ occurs that the next letter will be j. Th best guess for the next letter, when this $(N\text{-}1)$ gram is known to have occurred, will be that letter having the highest conditional probability. The second guess should be that with the second highest probability, etc. A machine or person guessing in the best way would guess letters in the orde of decreasing conditional probability. Thus the process of reducing a tex with such an ideal predictor consists of a mapping of the letters into th numbers from 1 to 27 in such a way that the most probable next lette [conditional on the known preceding $(N\text{-}1)$ gram] is mapped into 1, etc The frequency of 1's in the reduced text will then be given by

$$q_1^N = \Sigma p(i_1, i_2, \cdots, i_{N-1}, j) \tag{10}$$

where the sum is taken over all $(N\text{-}1)$grams $i_1, i_2, \bullet\bullet\bullet, i_{N-1}$ the j being tin one which maximizes p for that particular $(N\text{-}1)$ gram. Similarly, the frequency of 2's, q_2^N, is given by the same formula with j chosen to be tha letter having the second highest value of p, etc.

On the basis of N**-grams, a different set of probabilities for the** symbo in the reduced text, $q_1^{N+1}, q_2^{N+1}, \ldots, q_{27}^{N+1}$, would normally result. Since this prediction is on the basis of a greater knowledge of the past, one would expect the probabilities of low numbers to be greater, and in fact one can prove the following inequalities:

$$\sum_{i=1}^{S} q_i^{N+1} \geq \sum_{i=1}^{S} q_i^{N} \qquad S = 1, 2, \cdots . \tag{11}$$

This means that the probability of being right in the first S guesses when the preceding N letters are known is greater than or equal to that when only $(N\text{-}1)$ are known, for all S. To prove this, imagine the probabilities $p(i_1, i_2, \cdots, i_N > j)$ arranged in a table with j running horizontally and all the N-grams vertically. The table will therefore have 27 columns and 27^N rows. The term on the left of (11) is the sum of the S largest entries in each row summed over all the rows. The right-hand member of (11) is also a sum of entries from this table in which S entries are taken from each row but not necessarily the S largest. This follows from the fact that the right-hand member would be calculated from a similar table with $(N\text{-}1)$ grams rather than TV-grams listed vertically. Each row in the N-1 gram table is the sum of 27 rows of the N-gram table, since:

$$p(i_2, i_3, \cdots, i_N, j) = \sum_{i_1=1}^{27} p(i_1, i_2, \cdots, i_N, j). \tag{12}$$

The sum of the S largest entries in a row of the TV-1 gram table will equal the sum of the $27S$ selected entries from the corresponding 27 rows of the N-gram table only if the latter fall into S columns. For the equality in (11) to hold for a particular S, this must be true of every row of the TV-1 gram table. In this case, the first letter of the TV-gram does not affect the set of the S most probable choices for the next letter, although the ordering within the set may be affected. However, if the equality in (11) holds for all S, it follows that the ordering as well will be unaffected by the first letter of the N-gram. The reduced text obtained from an ideal N-1 gram predictor is then identical with that obtained from an ideal N-gram predictor.

Since the partial sums

$$Q_S^N = \sum_{i=1}^{S} q_i^N \qquad 5 = 1, 2, \cdots \tag{13}$$

are monotonic increasing functions of N, < 1 for all N, they must all approach limits as $N \rightarrow \infty$. Their first differences must therefore approach limits as $N \rightarrow \infty$, i.e., the q_i approach limits, q_i^∞. These may be interpreted as the relative frequency of correct first, second, $\cdots$, guesses with knowledge of the entire (infinite) past history of the text.

The ideal N-gram predictor can be considered, as has been pointed out, to be a transducer which operates on the language translating it into a sequence of numbers running from 1 to 27. As such it has the following two properties:

1. The output symbol is a function of the present input (the predicted next letter when we think of it as a predicting device) and the preceding (N-1) letters.
2. It is *instantaneously* reversible. The original input can be recovered by a suitable operation on the reduced text without loss of time. In fact, the inverse operation also operates on only the (N-1) preceding symbols of the reduced text together with the present output.

The above proof that the frequencies of output symbols with an N-gram predictor satisfy the inequalities:

$$\sum_1^S q_i^N > \sum_1^S q_i^{N-1} \qquad 5 = 1, 2, \cdots, 27 \qquad (14)$$

can be applied to any transducer having the two properties listed above, In fact we can imagine again an array with the various (N-1) grams listed vertically and the present input letter horizontally. Since the present output is a function of only these quantities there will be a definite output symbol which may be entered at the corresponding intersection of row and column, Furthermore, the instantaneous reversibility requires that no two entries in the same row be the same. Otherwise, there would be ambiguity between the two or more possible present input letters when reversing the translation. The total probability of the 5 most probable symbols in the output, say $\sum_1^S r_i$, will be the sum of the probabilities for5 entries in each row, summed over the rows, and consequently is certainly not greater than the sum of the S largest entries in each row. Thus we will have

$$\sum_1^S q_i^N \geq \sum_1^S r_i \qquad 5 = 1, 2, \cdots, 27 \qquad (15)$$

In other words ideal prediction as defined above enjoys a preferred position among all translating operations that may be applied to a language and which satisfy the two properties above. Roughly speaking, ideal prediction collapses the probabilities of various symbols to a small group more than any other translating operation involving the same number of letters which is instantaneously reversible.

Sets of numbers satisfying the inequalities (15) have been studied by Muirhead in connection with the theory of algebraic inequalities.[5] If (15) holds when the q_i^N and r_i are arranged in decreasing order of magnitude, and

[5] Hardy, Littlewood and Polya, "Inequalities," Cambridge University Press, 1934.

a $\sum_{1}^{27} q_i^N = \sum_{1}^{27} r_i$, (this is true here since the total probability in each case is 1), then the first set, q_i^N, is said to *majorize* the second set, r_i. It is known that the majorizing property is equivalent to either of the following properties:

1 The r_i can be obtained from the q_i by a finite series of "flows." By a flow is understood a transfer of probability from a larger q to a smaller one, as heat flows from hotter to cooler bodies but not in the reverse direction.

2 The r_i can be obtained from the q_i by a generalized "averaging" operation. There exists a set of non-negative real numbers, a_{ij}, with $\sum_j a_{ij} = \sum_i a_{ij} = 1$ and such that

$$r_i = \sum_j a_{ij}(q_j^N). \qquad (16)$$

5. ENTROPY BOUNDS FROM PREDICTION FREQUENCIES

If we know the frequencies of symbols in the reduced text with the ideal N-gram predictor, q_i, it is possible to set both upper and lower bounds to the N-gram entropy, F_N, of the original language. These bounds are as follows:

$$\sum_{i=1}^{27} i(q_i^N - q_{i+1}^N)\log i < F_N < -\sum_{i=1}^{27} q_i^N \log q_i^N. \qquad (17)$$

The upper bound follows immediately from the fact that the maximum possible entropy in a language with letter frequencies q_i^N is $-\sum q_i^N \log q_i^N$. Thus the entropy per symbol of the reduced text is not greater than this. The N-gram entropy of the reduced text is equal to that for the original language, as may be seen by an inspection of the definition (1) of F_N. The sums involved will contain precisely the same terms although, perhaps, in a different order. This upper bound is clearly valid, whether or not the prediction is ideal.

The lower bound is more difficult to establish. It is necessary to show that with any selection of N-gram probabilities $p(i_1, i_2, \ldots, i_N)$, we will have

$$\sum_{i=1}^{27} i(q_i^{\Lambda} - q_{i+1}^N)\log i < \sum_{i_1,\cdots,i_N} p(i, \cdots i_N)\log pi, \cdots i_{N-1}(i_N) \qquad (18)$$

The left-hand member of the inequality can be interpreted as follows: Imagine the q_i^N arranged as a sequence of lines of decreasing height (Fig. 3). The actual q_i^N can be considered as the sum of a set of rectangular distributions as shown. The left member of (18) is the entropy of this set of distributions. Thus, the i^{th} rectangular distribution has a total probability of

$i(q_i - q_{i+1}^N)$. The entropy of the distribution is log i. The total entropy then

$$\sum_{i=1} i(q_i^N - 9m) \log i.$$

The problem, then, is to show that any system of probabilities $p(i_1, \ldots i_N)$, with best prediction frequencies q_i has an entropy F_N greater than or equal to that of this rectangular system, derived from the same set of q_i.

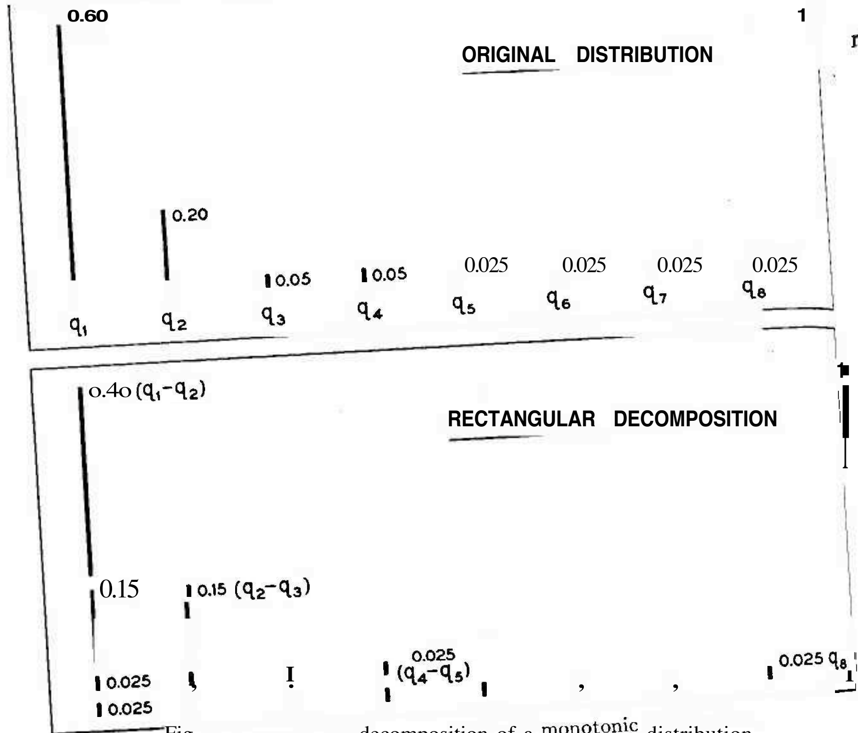

Fig. 3—Rectangular decomposition of a monotonic distribution.

The qi as we have said are obtained from the $p(i_1, \cdots, i_N)$ by arranging each row of the table in decreasing order of magnitude and adding vertically. Thus the q_i are the sum of a set of monotonic decreasing distributions. Replace each of these distributions by its rectangular decomposition. Each one is replaced then (in general) by 27 rectangular distributions; the q_i are the sum of 27 x 27^N rectangular distributions, of from 1 to 27 elements, and all starting at the left column. The entropy for this set is less than or equal to that of the original set of distributions since a termwise addition of two or more distributions always increases entropy. This is actually an application of the

general theorem that $H_y(x) < H(x)$ for any chance variables x and y. The equality holds only if the distributions being added are proportional. Now we may add the different components of the same width without changing the entropy (since in this case the distributions *are* proportional). The result is that we have arrived at the rectangular decomposition of the q_i, by a series of processes which decrease or leave constant the entropy, starting with the original N-gram probabilities. Consequently the entropy of the original system F_N is greater than or equal to that of the rectangular decomposition of the q_i. This proves the desired result.

It will be noted that the lower bound is definitely less than FN unless each row of the table has a rectangular distribution. This requires that for each

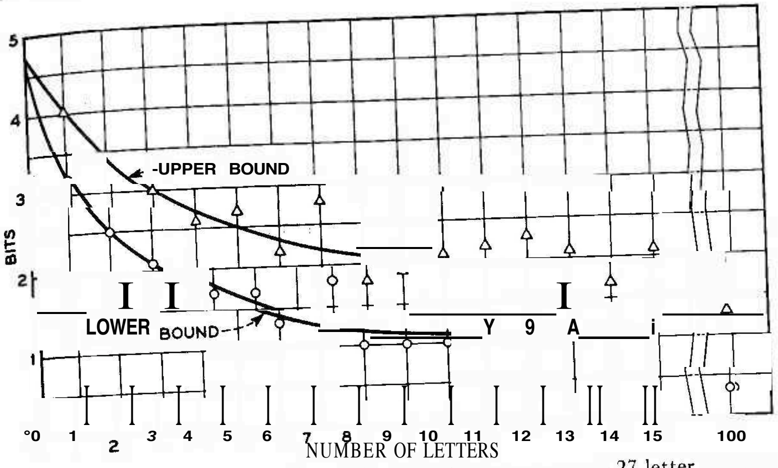

Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

possible $(N$-$1)$ gram there is a set of possible next letters each with equal probability, while all other next letters have zero probability.

It will now be shown that the upper and lower bounds for F_N given by (17) are monotonic decreasing functions of N. This is true of the upper bound since the q_i^{N+1} majorize the q_i^N and any equalizing flow in a set of probabilities increases the entropy. To prove that the lower bound is also monotonic decreasing we will show that the quantity

$$U = \sum_i i(qi - q_{i+1}) \log i \tag{20}$$

is increased by an equalizing flow among the q_i. Suppose a flow occurs from q_i to q_{i+1}, the first decreased by Δq and the latter increased by the same amount. Then three terms in the sum change and the change in U is given by

$$\Delta U = [-(i - 1) \log (i - 1) + 2i \log i - (i + 1) \log (i + 1)]\Delta q \tag{21}$$

The term in brackets has the form $-f(x-1) + 2f(x) - f(x+1)$ where $f(x) = x \log x$. Now $f(x)$ is a function which is concave upward for positive x, since $f''(x) = 1/x > 0$. The bracketed term is twice the difference between the ordinate of the curve at $x = i$ and the ordinate of the midpoint of the chord joining $i - 1$ and $i + 1$, and consequently is negative. Since Δq also is negative, the change in U brought about by the flow is positive. An even simpler calculation shows that this is also true for a flow from q_1 *to* q_2 or from q_{26} to q_{27} (where only two terms of the sum are affected). It follows that the lower bound based on the N-gram prediction frequencies q_i^N is greater than or equal to that calculated from the $N + 1$ gram frequencies q_i^{N+1}.

6. EXPERIMENTAL BOUNDS FOR ENGLISH

Working from the data of Table I, the upper and lower bounds were calculated from relations (17). The data were first smoothed somewhat to overcome the worst sampling fluctuations. The low numbers in this table are the least reliable and these were averaged together in groups. Thus, in column 4, the 47, 18 and 14 were not changed but the remaining group totaling 21 was divided uniformly over the rows from 4 to 20. The upper and lower bounds given by (17) were then calculated for each column giving the following results:

Column	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	100
Upper.	4.03	3.42	3.0	2.6	2.7	2.2	2.8	1.8	1.9	2.1	2.2	2.3	2.1	1.7	2.1	13
Lower.	3.19	2.50	2.1	1.7	1.7	1.3	1.8	1.0	1.0	1.0	1.3	1.3	1.2	.9	1.2	.6

It is evident that there is still considerable sampling error in these figures due to identifying the observed sample frequencies with the prediction probabilities. It must also be remembered that the lower bound was proved only for the ideal predictor, while the frequencies used here are from human prediction. Some rough calculations, however, indicate that the discrepancy between the actual F_N and the lower bound with ideal prediction (due to the failure to have rectangular distributions of conditional probability) more than compensates for the failure of human subjects to predict in the ideal manner. Thus we feel reasonably confident of both bounds apart from sampling errors. The values given above are plotted against N in Fig. 4.

ACKNOWLEDGMENT

The writer is indebted to Mrs. Mary E. Shannon and to Dr. B. M. Oliver for help with the experimental work and for a number of suggestions and criticisms concerning the theoretical aspects of this paper.