INFORMATION AND CONTROL 7, 304–306 (1964)

A Note on the Entropy of Words in Printed English

MARIO C. GRIGNETTI

Bolt Beranek and Newman Inc., Cambridge, Massachusetts

Shannon estimates the entropy of the set of words in printed English as 11.82 bits per word. As this figure seems inconsistent with some results deduced from several encoding procedures, the entropy was recalculated and found to be roughly 9.8 bits per word.

In one of his epoch-making papers, Shannon estimates the word-entropy of printed English as 11.82 bits per word. In obtaining this result, use was made of the Zipf formula, $p_n = .1/n$, relating word frequency to rank, as an approximation to the probability of the word. Bemer, however, using roughly the same approximation, calculates a mean length of 10.76 bits per word for an obviously nonoptimal compression code. As no apparent error was spotted in Bemer's derivation, this writer was led to inspect Shannon's result more closely.

Assuming with Shannon that Zipf's formula holds out to the n at which the total probability is unity, we have:

$$\sum_{n=1}^{N} = 10 \tag{1}$$

Recognizing in the left hand side of (1) a partial sum of the harmonic series, we can write, with Courant,

$$\sum_{n=1}^{N} \frac{1}{n} = \log N + C + E$$

where $C = 0.577216 \cdots$ is Euler-Mascheroni's constant and $E < 1/N$ is the error involved in the approximation. From (2) we get $N = 12370$ instead of $N = 8727$ as Shannon finds.

For the entropy H, expressing it in natural units, we have:

$$H = -\sum_{n=1}^{N} \frac{.1}{n} \log \frac{.1}{n} = \log 10 + .1 \sum_{n=1}^{N} \frac{\log n}{n} \tag{3}$$

The summation in the right hand side of (3) can be bounded by integrals,

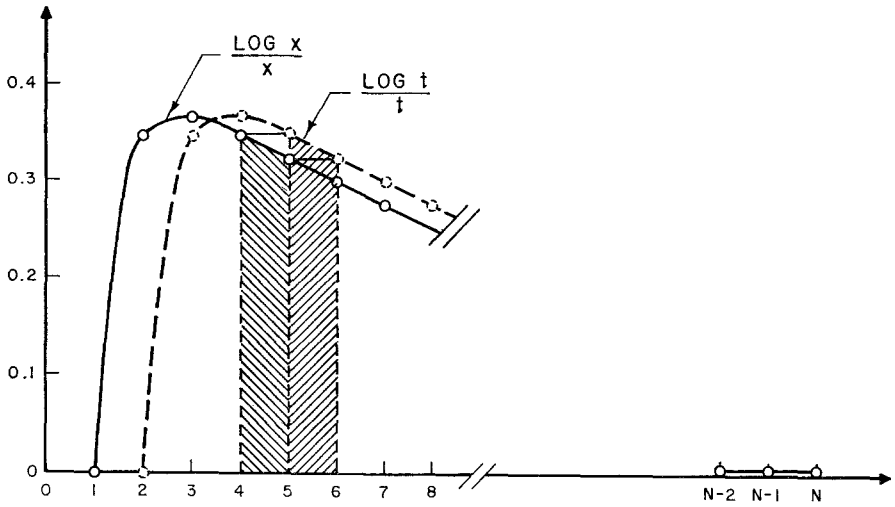

FIG. 1. To illustrate the derivation of Eqs. (4) and (5)

in the following way (see Fig. 1):

$$\int_3^N \frac{\log x}{x}\,dx < \sum_3^N \frac{\log n}{n} \tag{4}$$

whereas

$$\sum_4^N \frac{\log n}{n} < \int_3^{N-1} \frac{\log t}{t}\,dt \tag{5}$$

From (4) and (5) we obtain finally:

$$\int_3^N \frac{\log x}{x}\,dx + \frac{\log 2}{2} < \sum_1^N \frac{\log n}{n} < \int_3^{N-1} \frac{\log t}{t}\,dt + \frac{\log 2}{2} + \frac{\log 3}{3} \tag{6}$$

Evaluating the integrals, we have:

$$44.25 < \sum_1^N \frac{\log n}{n} < 44.62$$

and consequently, expressing H in bits,

$$9.72 < H < 9.76 \tag{7}$$

which differs from Shannon's result by more than two bits.

As these results rely so heavily on the Zipf's formula, which is only an approximation over a portion of the word frequency vs. rank curve and

fails particularly for very frequent words, an attempt was made to derive a better estimate of the entropy by computing it with the help of a word frequency table.

As all the word frequency tables known to this writer do not cover the very infrequent word range, the Zipf formula will still have to be used as an approximation over that portion of the distribution. According to Miller, this can be best done when the frequency table is extracted from a population containing about 120,000 words. Consequently, we have chosen Dewey's tables, drawn over a sample of more than 100,000 words and containing 10,119 different words as the basis for our computation.

The portion of the entropy corresponding to the tabulated words (the commonest 1027 words, representing 78.6% of the total usage) was computed with the help of a digital computer and was found to be 6.775 bits.

For the remaining 9092 words an artificial, discrete distribution was constructed so that (a) it accounted for an additional 20.3% of the usage, (b) it followed closely Zipf's law, and (c) was consistent with the tabulated distribution.

This gives an additional 3.052 bits towards the total value of the entropy. As the remaining 1.1% of the word frequency distributions is not likely to affect the total value beyond the second decimal figure we can safely assume $H = 9.83$ bits as an estimate to the entropy of words in printed English.

RECEIVED: October 4, 1963

REFERENCES

BEMER, R. W. (Oct. 1960), Do it by the numbers—Digital Shorthand. *Commun. Assoc. Comput. Mach.* **3,** No. 10, 530–536.

COURANT, R. (1937), "Differential and Integral Calculus," 2nd ed. Blackie, Glasgow.

DEWEY, G. (1950), "Relative Frequency of English Speech Sounds." Harvard Univ. Press, Cambridge, Massachusetts.

MILLER, G. A. (1951), "Language and Communication." McGraw-Hill, New York.

SHANNON, C. E. (1951), Prediction and entropy of printed English. *Bell System Tech. J.* **30,** No. 1, 50–64.

ZIPF, G. K. (1949), "Human Behavior and the Principle of Least Effort." Addison-Wesley, Cambridge, Massachusetts.