AD__________ Accession No.__________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED

AD__________ Accession No.__________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED

AD__________ Accession No.__________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED

AD__________ Accession No.__________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED

2144-147-T

Frequency of Occurrence of Words

A Study of Zipf's Law, with Application to Mechanical Translation

Andreas M. Koutsoudas and Robert E. Machol
(With an Appendix by George J. Minty)

June 1957

The University of Michigan
Engineering Research Institute
Willow Run Laboratories
Willow Run Airport
Ypsilanti, Michigan

ABSTRACT

Existing laws concerning the frequencies of words in language — specifically Zipf's and Joos' laws — are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

Project MICHIGAN is a research and development project which has been carried on by The University of Michigan since May of 1953 under a tri-service charter administered by the U. S. Army Signal Corps. The objective of the project is the continual improvement of the capabilities of the armed forces for battle-area surveillance, the mission of which is to supply field commanders and their staffs, at all echelons, with information necessary for the effective utilization of all combat weapons through planning and tactical decisions.

The objective of the project is accomplished by research and development work on sensory and data-processing devices and techniques; integration of these and other surveillance devices into a continually improving battle-area surveillance system; and, as appropriate, advising the military on matters of research, development, and procurement within the field of battle-area surveillance.

Activities of the project include the development of basic system concepts and designs, the improvement of such designs by the continuing integration of improved subsystems, the design and specification of subsystems and their evaluation by field test and simulation, the development of sensory devices and data-processing techniques, and the review of research programs pertinent to battle-area surveillance. The work on the development of sensory devices includes basic and applied research in the fields of optics and vision, acoustics and seismics, radar, and infrared.

The University of Michigan, late in 1955, began a program of research to investigate the possibility of mechanically translating language. The ultimate goal of this research is to develop a process by which a foreign text in a particular field can be translated into precise and unambiguous English without the intervention of a human pre-editor or post-editor. Indispensible to this process is the provision of the following four factors: (1) a dictionary of words, (2) a set of rules for the identification of the words, (3) a set of rules to handle the syntactical function of the words, and (4) a set of rules to differentiate between multiple meanings of the words. All four of these factors must be capable of being stated in such a way that they can be easily acted upon by an electronic computer. These factors have been previously discussed[1]; we shall limit ourselves to the first factor, namely that of the dictionary.

The immediate problem arising with regard to the compilation of a dictionary is mostly an economic one. Can a dictionary be compiled, for instance, which on the one hand would be large enough to represent accurately the language universe in mind, and on the other, be small enough to be contained within an electronic memory? Specifically, can it be safely stated, for example, that the number of different words (lexical or otherwise) needed to translate Russian astronomy papers into correct and unambiguous English is small enough so as not to exceed an electronic memory for a given period of time?

We introduce the following model of a language. The author is considered to be a population (in the classical statistical sense) who generates samples of words of size n. Each word in the sample is assumed to be independent of the preceding words (in fact, of course, there is some autocorrelation, but since this drops off to zero rather rapidly[2] the assumption of independence will have little effect on large samples). The total number of different words in the population is M; this concept is ill-defined, since an author may learn new words, coin neologisms, and the like, but the assumption that M is a constant is also a useful and reasonable one for a first approximation. This model is discussed further below.

For the present purpose, two words will be considered the same if they are spelled alike, and different if they are spelled differently. Thus "bridge" and "bridges" will be considered different words, but "bridge" (a structure) and "bridge" (a game) will be considered the same word, as will "bridge" (noun) and "bridge" (verb). For the dictionary to be constructed in developing mechanical translation, somewhat different definitions of a word will be necessary, but the present assumption is most useful for using machinery to make counts of words in large samples of text.

Each word in the population has a certain probability, P, of being generated. Each of the M words is to be given a rank, R, $1 \leqslant R \leqslant M$, such that $P(R_1) \geqslant P(R_2)$ for $R_2 > R_1$. It is clearly a necessary condition that

$$\sum_{R=1}^{M} P(R) = 1. \qquad (1)$$

One of the problems at hand is to estimate M from an analysis of a sample of text. For example, from an analysis of Russian astronomy articles, it is desired to estimate the total number of different words which might be used by Russian astronomers. More specifically, it is desired to estimate how frequently a word will be found which is not in the dictionary if a dictionary of given size is compiled.

The first serious effort to solve this problem was made by Zipf[3] in connection with a series of sociological investigations. He was able to show that a number of factors could be related by the formula $P(R) = a/R^c$, where R is the rank (as defined above) and a and c are constants to be determined experimentally. He found that such formulas applied to things as different as the populations of cities (with c=1) and the incomes of people (with c=1/2), and he also applied it in our present context, asserting that using the constants a = 1/10 and c = 1 gave a good fit to the data.

It was obvious even from Zipf's small sample of data that this formula did not give a good fit for small values of R; for example, subsequent counts of large samples have shown that in English $P(1) \approx 0.07$, rather than 0.10 as predicted by Zipf's law. A more serious difficulty is that there was no adequate method of testing the formula for large R. As a typical example, in the 30,000 words[5] of Pushkin's "The Captain's Daughter," half of all the words occurred only once, and half of the remainder occurred only two or three times; there is no method of estimating the true value of P(R) for these words, which are precisely the words of greatest interest for our purposes.

However, the product of R and P(R) for most of the intermediate words in Zipf's small samples was approximately 1/10, and this sufficed to demonstrate Zipf's basic premise: that there were certain fundamental underlying principles controlling such sociological factors as the use of words, and that these principles were similar over a wide variety of sociological phenomena.

Very little may be deduced from the empirical observation that the products $R \times P(R)$ are constant over a comparatively small range of R, because it is difficult to conceive of a structure of language which would not lead to such an observation. The index R is constrained to go up slowly and with equal increments by definition, and the probability P(R) is constrained to go down slowly and with more or less equal increments by statistical considerations. Finally, the constraint of (1) forces these increments to balance one another approximately, so that the products $R \times P(R)$ must remain approximately constant over a small range of R.

Zipf's law may be written in the form

$$P(R) = \frac{1}{10}\,\frac{1}{R}. \tag{2}$$

Substituting (2) into (1) we obtain

$$\sum_{R=1}^{M} \frac{1}{R} = 10, \tag{3}$$

where the quantity under the summation sign is the well known harmonic series, $1 + 1/2 + 1/3 + 1/4 + \ldots$ whose sum is approximately $0.5772 + \log_e M$. (In fact, Zipf's law is often called the "harmonic series law").

Zipf did not comment on the magnitude of M, but presumably he tacitly assumed that it was infinite; such an assumption is, of course, incompatible with (3). If, however, (3) is modified to the following form,

$$\frac{1}{10}\sum_{R=1}^{M} \frac{1}{R^{1+b}} = 1, \tag{4}$$

the series becomes convergent, so long as b is any positive number, no matter how small. Thus, we may replace (2) by

$$P(R) = \frac{1}{10}\,\frac{1}{R^{1+b}}. \tag{5}$$

Equation (5) is as good a fit to the experimental data (which are very meagre) as is (2), and for any value of M a value of b can be found which makes (5) hold. These facts, as applied to Zipf's law, were first pointed out by Joos[4], and the constant b, in this context is known as Joos' constant. Thus, for $M = \infty$, $b \approx 0.106$. For $b = 0$, $M \approx 12{,}000$.

However, the fact that a value of b can be found which will make (4) consistent for any value of M is by no means a proof that (5) is in accordance with the facts. To substantiate (5) by the methods of Zipf and Joos would require extremely large samples, in which all the words were ranked and counted so that the R's could be compared with the P(R)'s over a large range of R, and some sort of statistical test applied to a comparison of the results with the law. Such a procedure would be excessively tedious, and has not been performed.

The difficulty, as pointed out before, is that the most interesting part of the sample is the words which appear most infrequently; and unless the sample is enormous (say many millions of words) it is difficult to obtain any statistical significance from the frequency of appearance of these words. However, there are other statistics of the sample which are more sensitive to the parameters which we are investigating. The present work was therefore an attempt to find such a statistic; that is, to derive new mathematical formulas by which these laws, (2) and (5), and others like them, might be tested against easily obtainable experimental data.

General formulas for two such statistics have been derived. One is $V = V(n)$, the number of different words in a sample of n words; the other is $U = U(n)$, the number of words which appear only once in a sample of n words. Each of these statistics is sensitive to the behavior of P(R) for large R, as desired. It should be noted that

$$\lim_{n\to\infty} V = M \tag{6}$$

always holds, and may in fact serve as a definition of M. Also,

$$\lim_{n\to\infty} U = 0$$

holds if M is finite.

In general, given the distribution function of R — that is, given P(R) as a function of R — it is possible to write the distribution function of V —

that is, to write P(V) as a function of n and M. This distribution function is excessively complex, but we can write its mean easily, as follows. The probability that a particular word is the word of rank R is

$$P(R),$$

and the probability that it is not the word of rank R is

$$1 - P(R).$$

Then the probability that no one of a sample of n words will be the word of rank R is

$$[1 - P(R)]^n,$$

and the probability that the word of rank R will appear at least once in the sample of n words is

$$1 - [1 - P(R)]^n.$$

Then the expected value of V is simply the sum of these probabilities over all values of R:

$$E(V) = \sum_{R=1}^{M} \left\{ 1 - [1 - P(R)]^n \right\} \qquad (7)$$

$$= M - \sum_{R=1}^{M} [1 - P(R)]^n.$$

Substituting (5) into (7), we obtain

$$E(V) = \sum_{R=1}^{M} \left[1 - \left(1 - \frac{1}{10R^{1+b}}\right)^n \right] \qquad (8)$$

$$= M - \sum_{R=1}^{M} \left(1 - \frac{1}{10R^{1+b}}\right)^n$$

subject to the constraint of (4), which defines b as a function of M. Equation (8) reduces to (4) for n = 1, and tends asymptotically to M for large n.

The numerical evaluation of (8) is tedious, and is given in the appendix. The result is shown in Figure 1, where the expected value of V is plotted as a function of n for several values of the parameter M. Also shown on the graph are several points taken from Josselson's word counts[5] of a novel by Pushkin.

It is evident from Figure 1 that (8), and therefore (5), are not adequate descriptions of the data. If a sufficiently small value of M (about 5600) is taken to make the curve pass through the first point, then M is so small as to be intuitively unacceptable; furthermore, the curve then passes well below the points for larger n. Curves for large values of M rise to large values of V(n) considerably more rapidly than do the experimental data.

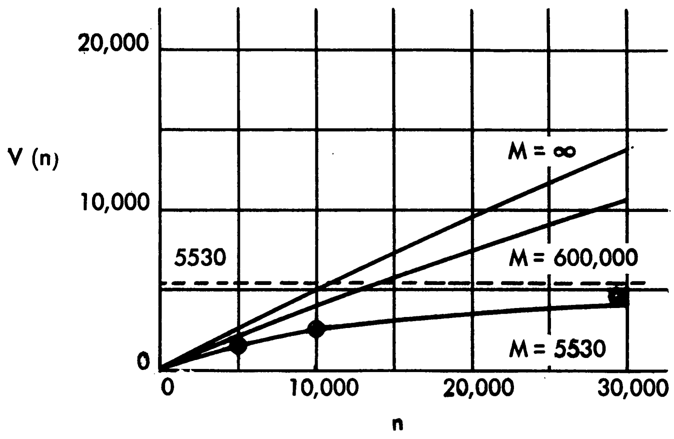

FIG 1

In other words, the difference $V(n_1) - V(n_2)$, if n_1 and n_2 are large numbers, is larger for actual language than is predicted by Zipf's law or Joos' modification of this law. This difference represents the number of different words which will be found in a sample of size n_2 and which will not be found in the dictionary if the dictionary has been made up to include only those words found in a portion n_1 of the sample. (The total number of words in n_2 which are not in the dictionary is given by

$$\frac{dV}{dn}(n_2 - n_1),$$

where dV/dn is evaluated at n_1; this quantity is greater than $V(n_1) - V(n_2)$ because some of the new words will be repeated.)

The derivation of U is similar. The probability that the first word in the sample is the word of rank R is

$$P(R)$$

and the probability that all the remaining words in the sample are other words than the word of rank R is

$$[1 - P(R)]^{n-1}.$$

Then the joint probability of these occurrences — i.e., the probability that the word of rank R shall appear just once, and that it shall be the first word — is

$$P(R)\,[1 - P(R)]^{n-1}.$$

But the probability that the word of rank R shall occur just once in some other position than the first is exactly the same; and there are just n such possibilities. Hence, the probability that the word of rank R shall occur just once in the sample is

$$n\, P(R) \left[1 - P(R)\right]^{n-1},$$

and the expected value of U is

$$E(U) = n \sum_{R=1}^{M} P(R) \left[1 - P(R)\right]^{n-1}, \qquad (9)$$

Equation (9) holds no particular advantages over (7) for purposes of numerical evaluation. However, it is possible to derive one interesting result. Dividing (9) by n, and subtracting $\sum (1-P)^{n-1}$ from both sides, we obtain

$$\frac{E(U)}{n} - \sum_{R=1}^{M} \left[1 - P(R)\right]^{n-1} = \sum_{R=1}^{M} \left[P(R) - 1\right]\left[1 - P(R)\right]^{n-1}$$

$$= - \sum_{R=1}^{M} \left[1 - P(R)\right]^{n}.$$

Hence,

$$\frac{E(U)}{n} - E\left[M - V(n-1)\right] = - E\left[M - V(n)\right]$$

or

$$\frac{E(U)}{n} = E\left[V(n) - V(n-1)\right] \approx \frac{dV}{dn}. \qquad (10)$$

The approximation involved in deriving (10) consists only of substituting dV/dn for V(n) — V(n-1), which involves negligible error for large n; the substitution of V for E(V), the error of which depends on the variance of the data; and the assumption of independence, which is discussed below. Thus, (10) may be useful in two ways. It may be used to estimate dV/dn, which, as mentioned above, defines the number of words which will be found in a new sample of text which are not in the dictionary (if it turns out to be simpler to count U for the sample from which the dictionary was made than to count various values of $V(n_i)$ and compute dV/dn directly); or it may be used to give an independent estimate of the variance of the data, by computing dV/dn by both methods. This will make it possible to place confidence limits around the prediction of the frequency with which words not in the dictionary will be found in new samples of text from the same population.

The emphasis on the same population is a necessary one. Different samples of text from different authors may differ in at least three ways (Fig. 2):

(1) They may have different values of M (thus, James Joyce had an unusually large vocabulary, as is evident from the word count[6] of "Ulysses") and the value of M will vary with our definition of a word, with the amount of inflection in the language, and so forth.

(2) For the same value of M, the words in the vocabulary may be different; in particular, over and above the ordinary words there will be certain technical words in any special usage.

(3) The language may have different structures, in the sense that the relationship between R and P(R) may vary. In fact, there is very little evidence to verify the assumption, tacitly made by Zipf, that this structure is constant from one corpus to another.

While the data of Figure 2 are not strictly commensurable (thus, in the Russian word counts, different inflectional forms were generally counted as one word, in the "Representative Modern English" they were counted as different words, and in the correspondence the counting rules are not known), it is clear that they do not come from a single population.

While we do not have enough data to find the empirical relationships between R and P(R), and the science of linguistics is not far enough advanced to predict them theoretically, consideration of the nature of the problem indicates that it is probably very complex. For example, although Shannon[2] originally asserted that the long-range structure of English was comparatively unimportant, in his later studies[7] he has shown that the opposite is true. From his point of view, this means that the entropy of English is less than half of the 2.5 bits per symbol which he originally estimated; from our point of view it means that the independence of successive words which we assumed may not be justified. Thus if, in the first ten words of a 10,000 word text, we find the word "cathode," we are much more likely

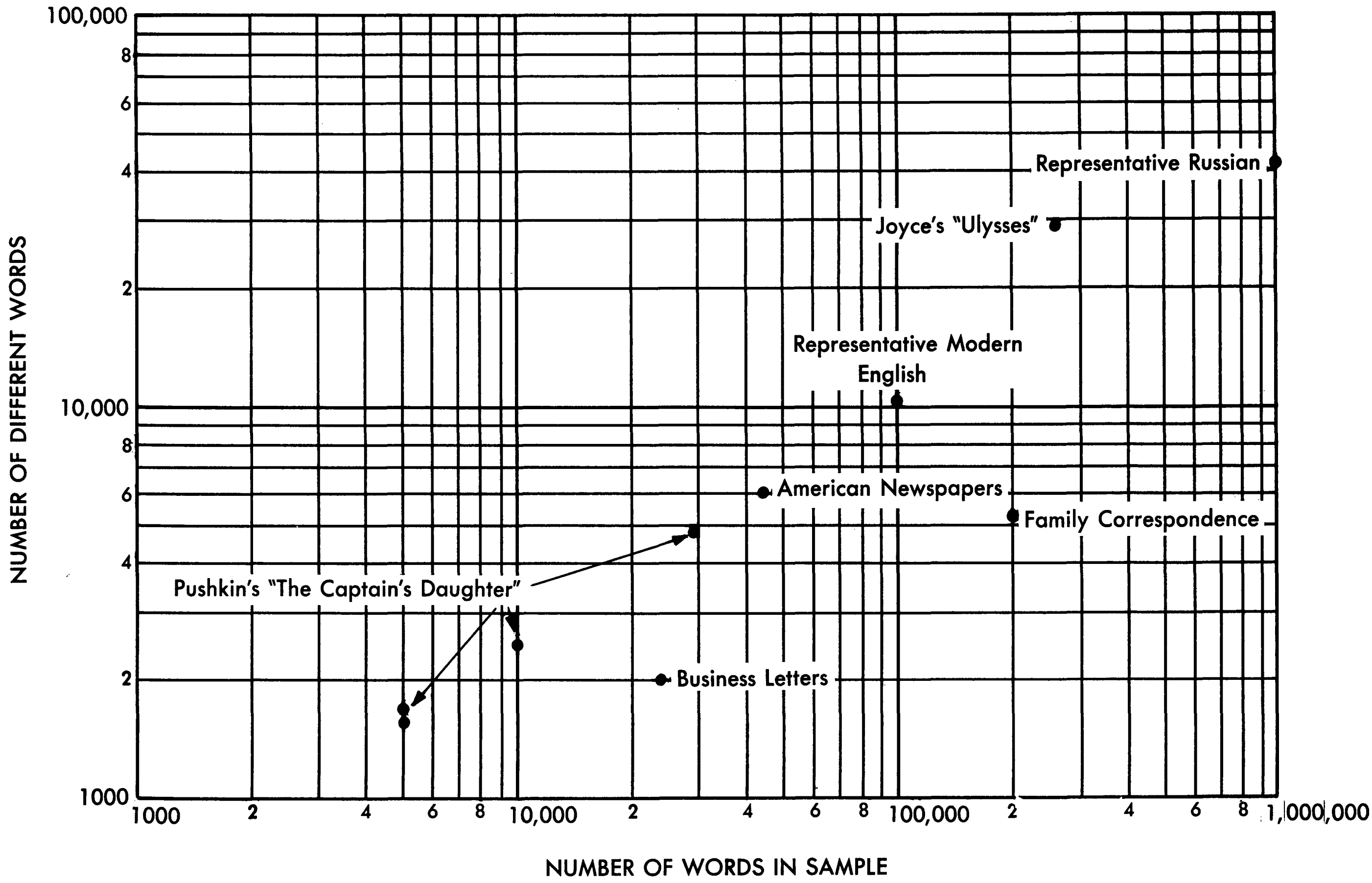

FIG. 2

to find the word "amplifier," even near the end of the sample, than the word "tablespoonful." In part, this is taken care of by restricting ourselves to a given type of text, such as Russian astronomy articles, and assuming that this has a fixed vocabulary, of size M, which includes such words as "galaxy" with comparatively high probability and such words as "tablespoonful" with comparatively low probability. Even within this restriction, however, there is a high degree of correlation over the span of, say, a 3000-word article, so that the word "galaxy" or "calcium" or "corona" might appear many times in one article and not at all in another. These comments are adduced merely to indicate that the structure of language is exceedingly complex, and is not likely to be greatly illuminated by unsophisticated approaches or analyses of small samples of data.

From the point of view of mechanical translation, therefore, it appears that an empirical approach will be the most valuable in attempting to define the necessary dictionary size, and the frequency of words which will appear in the text but which are not in a dictionary of given size. This empirical approach, consisting primarily of making counts of V(n) for samples of very large size, may be supplemented by formulas such as (10) which are not dependent on any particular formula for the probability of occurrence of particular words.

APPENDIX A

George J. Minty

Consideration of $M = \infty$; $M = 600,000$

$V(n) \approx n$ for small n, and $V(n) \approx M$ for very large n (if M is finite). We have now to derive an approximation-formula valid for intermediate value of n, and for large n if M is infinite. We have succeeded in this task in the case $M > 12,000$, $b > 0$.

Consider

$$V(n) = \sum_{R=1}^{M} \left[1 - \left(1 - \frac{1}{aR^{\alpha}} \right)^n \right],$$

where $a > 1$

$$\alpha = 1 + b > 1.$$
$$M > 1.$$

It can be written as

$$V(n) = \int_{R=1}^{M} \left[1 - \left(1 - \frac{1}{aR^{\alpha}} \right)^n \right] dR + E_1,$$

where $0 < E_1 < 1 - \left(1 - \frac{1}{a}\right)^n < 1$.

For the sum is the sum of the areas of the rectangles illustrated in Figure A-1; the integral is the area under the curve; their difference is the sum of the areas of the shaded "triangles," which, by translation to the left, can all be fitted without overlapping into the first rectangle, whose area is < 1.

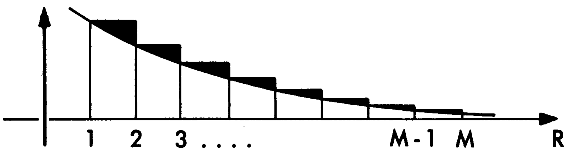

FIG. A-1

Now we make a change of variables in the integral. Let $x = \frac{n}{aR^{\alpha}}$.

So $R = \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} x^{-\frac{1}{\alpha}}$, and

$$dR = -\frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} x^{-\frac{1}{\alpha}-1} dx.$$

Then

$$V(n) = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{aM^{\alpha}}}^{\frac{n}{a}} \left[1 - \left(1 - \frac{x}{n}\right)^n\right] x^{-\frac{1}{\alpha}-1} dx + E_1,$$

$$\text{or } V(n) = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{aM^{\alpha}}}^{\frac{n}{a}} \left[1 - e^{-x}\right] x^{-\frac{1}{\alpha}-1} dx + E_1 + E_2 ,$$

where

$$E_2 = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{aM^{\alpha}}}^{\frac{n}{a}} \left[e^{-x} - \left(1 - \frac{x}{n}\right)^n\right] x^{-\frac{1}{\alpha}-1} dx .$$

$$E_2 = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{aM^{\alpha}}}^{\frac{n}{a}} \left[1 - \frac{\left(1 - \frac{x}{n}\right)^n}{e^{-x}}\right] e^{-x} x^{-\frac{1}{\alpha}-1} dx .$$

$$\text{Consider } \ln \frac{\left(1 - \frac{x}{n}\right)^n}{e^{-x}} = n \ln \left(1 - \frac{x}{n}\right) - \ln e^{-x} .$$

In the range of integration, $x \leqslant \frac{n}{a}$, so $\frac{x}{n} \leqslant \frac{1}{a} < 1$ so the logarithm can be expanded in a power series:

$$= n\left(- \frac{x}{n} - \frac{x^2}{2n^2} - \frac{x^3}{3n^3} - \ldots\right) + x$$

$$= nx^2\left(- \frac{1}{2n^2} - \frac{x}{3n^3} - \ldots\right) .$$

This expression is $\leqslant 0$, so $\left(1 - \frac{x}{n}\right)^n \leqslant e^{-x}$

and $e^{-x} - \left(1 - \frac{x}{n}\right)^n \geqslant 0$, so E_2 is clearly positive.

Now, $x \leqslant \frac{n}{a}$, so

$$\ln \frac{\left(1 - \frac{x}{n}\right)^n}{e^{-x}} \geqslant nx^2\left(- \frac{1}{2n^2} - \frac{\left(\frac{n}{a}\right)}{3n^3} - \frac{\left(\frac{n}{a}\right)^2}{4n^4} - \ldots\right)$$

$$\geqslant \frac{a^2x^2}{n}\left(- \frac{\left(\frac{n}{a}\right)^2}{2n^2} - \frac{\left(\frac{n}{a}\right)^3}{3n^3} - \frac{\left(\frac{n}{a}\right)^4}{4n^4} - \ldots\right)$$

$$\geqslant \frac{a^2x^2}{n}\left(-\frac{1}{2a^2}-\frac{1}{3a^3}-\frac{1}{4a^4}\cdots\right)$$

$$\geqslant \frac{a^2x^2}{n}\left(-\frac{1}{a}-\frac{1}{2a^2}-\frac{1}{3a^3}-\frac{1}{4a^4}+\frac{1}{a}\right)$$

$$\geqslant \frac{a^2x^2}{n}\left(\ell n\left(1-\frac{1}{a}\right)+\frac{1}{a}\right).$$

Note that the bracketed expression is negative:

$$\geqslant -\frac{ax^2}{n}\left|1+a\ell n\left(1-\frac{1}{a}\right)\right|$$

Since the logarithm is a monotonic function, we have

$$\frac{\left(1-\frac{x}{n}\right)^n}{e^{-x}} \geqslant e^{-\frac{ax^2}{n}\left|1+a\ell n\left(1-\frac{1}{a}\right)\right|}$$

$$1-\frac{\left(1-\frac{x}{n}\right)^n}{e^{-x}} \leqslant 1-e^{-\frac{ax^2}{n}\left|1+a\ell n\left(1-\frac{1}{a}\right)\right|}.$$

Examining the function $1 - e^{-z} - z$ for $0 \leqslant z < \infty$, we see that it is zero for $z = 0$, and its derivative is $e^{-z} - 1$, which is $\leqslant 0$. So the function is $\leqslant 0$ in this range:

$$1 - e^{-z} - z \leqslant 0.$$

$$1 - e^{-z} \leqslant z\ .$$

So we conclude

$$1-\frac{\left(1-\frac{x}{n}\right)^n}{e^{-x}} \leqslant \frac{ax^2}{n}\left|1+a\ell n\left(1-\frac{1}{a}\right)\right|,$$

and therefore

$$E_2 \leqslant \frac{a}{n\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}}\left|1+a\ell n\left(1-\frac{1}{a}\right)\right|\int_{\frac{n}{aM^\alpha}}^{\frac{n}{a}} xe^{-x}\,x^{-\frac{1}{\alpha}}\,dx.$$

Now, $xe^{-x} \geqslant 0$ for $x \geqslant 0$. It is zero for $x = 0$ and for $x \to \infty$. Its maximum is found by setting the derivative = 0:

$$-xe^{-x} + e^{-x} = 0$$

$$e^{-x}(x - 1) = 0$$

$$x - 1 = 0$$

$$x = 1$$

So the value of the maximum is $1 \cdot e^{-1} = \frac{1}{e}$. Also we change the lower limit of the integral to zero:

$$E_2 \leqslant \frac{a}{n\alpha e}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \left|1 + a\ln\left(1 - \frac{1}{a}\right)\right| \int_0^{\frac{n}{a}} x^{-\frac{1}{\alpha}}\, dx$$

$$\leqslant \frac{a}{n\alpha e}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \left|1 + a\ln\left(1 - \frac{1}{a}\right)\right| \frac{\left(\frac{n}{a}\right)^{1-\frac{1}{\alpha}}}{1 - \frac{1}{\alpha}}$$

$$\leqslant \frac{\alpha}{(\alpha - 1)\,\alpha e} \cdot \left|1 + a\ln\left(1 - \frac{1}{a}\right)\right|$$

$$\leqslant \frac{\left|1 + a\ln\left(1 - \frac{1}{a}\right)\right|}{(\alpha - 1)\, e}.$$

Now we go back to V (n).

$$V(n) = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{aM^{\alpha}}}^{\frac{n}{a}} \left[1 - e^{-x}\right] x^{-\frac{1}{\alpha}-1}\, dx + E_1 + E_2 = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{aM^{\alpha}}}^{\infty} \left[1 - e^{-x}\right] x^{-\frac{1}{\alpha}-1}\, dx + E_1 + E_2 - E_3,$$

where

$$E_3 = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{a}}^{\infty} \left[1 - e^{-x}\right] x^{-\frac{1}{\alpha}-1}\, dx.$$

Now, $1 - e^{-x}$ is less than 1; so

$$E_3 \leqslant \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{a}}^{\infty} x^{-\frac{1}{\alpha}-1}\, dx$$

$$\leqslant \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \left. \frac{x^{-\frac{1}{\alpha}}}{-\frac{1}{\alpha}} \right]_{\frac{n}{a}}^{\infty}$$

$$\leqslant \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \left(\frac{n}{a}\right)^{-\frac{1}{\alpha}} = 1.$$

So now we have

$$V(n) = \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_{\frac{n}{aM^{\alpha}}}^{\infty} (1 - e^{-x})\, x^{-\frac{1}{\alpha}-1}\, dx + E_1 + E_2 - E_3$$

$$= -\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_0^{\infty} (1 - e^{-x})\, d\left(x^{-\frac{1}{\alpha}}\right) + E_1 + E_2 - E_3 - \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_0^{\frac{n}{aM^{\alpha}}} (1 - e^{-x})\, x^{-\frac{1}{\alpha}-1}\, dx.$$

Integrating by parts

$$V(n) = -\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \left. (1 - e^{-x})\, x^{-\frac{1}{\alpha}} \right]_0^{\infty} + \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_0^{\infty} x^{-\frac{1}{\alpha}} e^{-x}\, dx$$

$$+ E_1 + E_2 - E_3 - \frac{1}{\alpha}\left(\frac{n}{\alpha}\right)^{\frac{1}{\alpha}} \int_0^{\frac{n}{aM^{\alpha}}} (1 - e^{-x})\, x^{-\frac{1}{\alpha}-1}\, dx.$$

The product $(1 - e^{-x})\, x^{-\frac{1}{\alpha}}$ approaches zero as $x \to \infty$; as $x \to 0$ we use l'Hôpital's rule:

$$\lim_{x \to 0} \frac{1 - e^{-x}}{\frac{1}{x^{\alpha}}} = \lim_{x \to 0} \frac{e^{-x}}{\frac{1}{\alpha}\frac{1}{x^{\alpha}} - 1} = \alpha \lim_{x \to 0} x^{1 - \frac{1}{\alpha}} e^{-x} = 0.$$

So

$$V(n) = \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_0^{\infty} x^{-\frac{1}{\alpha}} e^{-x}\, dx + E_1 + E_2 - E_3 - \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_0^{\frac{n}{aM^{\alpha}}} (1 - e^{-x})\, x^{-\frac{1}{\alpha} - 1}\, dx$$

$$= \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \Gamma\left(1 - \frac{1}{\alpha}\right) + E_1 + E_2 - E_3 - \frac{1}{\alpha}\left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \int_0^{\frac{n}{aM^{\alpha}}} (1 - e^{-x})\, x^{-\frac{1}{\alpha} - 1}\, dx,$$

where $E_1 < 1$; $E_2 < \frac{1 + 10\,\ln \cdot 9}{b \cdot 2.71} = \frac{0.2}{b}$; $E_3 < 1$ (b small)

For $M = \infty$ or $M = 600,000$, the total error is less than 3.

For $M = \infty$, the latter integral is zero.

For M finite and very large, the integral can be evaluated by use of power-series:

$$\int_0^{\frac{n}{aM^{\alpha}}} (1 - e^{-x})\, x^{-\frac{1}{\alpha} - 1}\, dx$$

$$= \int_0^{\frac{n}{aM^{\alpha}}} \left(x - \frac{x^2}{2!} + \frac{x^3}{3!} - \dots\right) x^{-\frac{1}{\alpha} - 1}\, dx \int_0^{\frac{n}{aM^{\alpha}}} \left(x^{-\frac{1}{\alpha}} - \frac{x^{1 - \frac{1}{\alpha}}}{2!} + \frac{x^{2 - \frac{1}{\alpha}}}{3!} - \dots\right) dx$$

$$= \left[\frac{x^{1 - \frac{1}{\alpha}}}{1 - \frac{1}{\alpha}} - \frac{x^{2 - \frac{1}{\alpha}}}{\left(2 - \frac{1}{\alpha}\right)2!} + \frac{x^{3 - \frac{1}{\alpha}}}{\left(3 - \frac{1}{\alpha}\right)3!} - \dots\right]_0^{\frac{n}{aM^{\alpha}}}$$

$$= \left[\alpha\, x^{-\frac{1}{\alpha}} \left(\frac{x}{\alpha - 1} - \frac{x^2}{(2\alpha - 1)\, 2!} + \frac{x^3}{(3\alpha - 1)\, 3!} - \dots\right)\right]_0^{\frac{n}{aM^{\alpha}}}$$

$$= \left(\alpha \frac{n}{aM^{\alpha}}\right)^{-\frac{1}{\alpha}} \left[\frac{\left(\frac{n}{aM^{\alpha}}\right)}{\alpha - 1} - \frac{\left(\frac{n}{aM^{\alpha}}\right)^2}{(2\alpha - 1)\, 2!} + \frac{\left(\frac{n}{aM^{\alpha}}\right)^3}{(3\alpha - 1)\, 3!} - \dots\right].$$

So for finite M, we can use

$$V(n) = \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \Gamma\left(1 - \frac{1}{\alpha}\right) + E_1 + E_2 - E_3 - \frac{M\left(\frac{n}{aM^{\alpha}}\right)}{\alpha - 1} + \frac{M\left(\frac{n}{aM^{\alpha}}\right)^2}{(2\alpha - 1) \cdot 2!} - \frac{M\left(\frac{n}{aM^{\alpha}}\right)^3}{(3\alpha - 1) \cdot 3!} + \ldots$$

If $n < aM^{\alpha}$, the error introduced by using only a few terms of the series is less than the first term dropped off. For example, if $n = 10,000$, $a = 10$, $M = 600,000$, $\alpha = 1.055$, the third term of the series is about

$$\frac{100,000 \times \left(\frac{1}{700}\right)^3}{2.3} \doteq \frac{1}{7000}$$

so that the error introduced by dropping all but the first two terms will be much smaller than the errors E_1, E_2, E_3.

To read $\Gamma\left(1 - \frac{1}{\alpha}\right) = \Gamma\left(\frac{b}{1 + b}\right)$

from tables of $\Gamma(x)$ $(1 \lesssim x \lesssim 2)$

use the formula $\Gamma\left(\frac{b}{1 + b}\right) = \frac{1 + b}{b} \Gamma\left(\frac{1 + 2b}{1 + b}\right)$.

We have thus derived the formulas

$$V(n) \doteq \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \Gamma\left(1 - \frac{1}{\alpha}\right)$$

for M infinite, and

$$V(n) \doteq \left(\frac{n}{a}\right)^{\frac{1}{\alpha}} \Gamma\left(1 - \frac{1}{\alpha}\right) - \frac{M\left(\frac{n}{aM^{\alpha}}\right)}{\alpha - 1} + \frac{M\left(\frac{n}{aM^{\alpha}}\right)^2}{(2\alpha - 1) \cdot 2!}$$

for M finite and very large, and for $n < a M^{\alpha}$.

REFERENCES

1. Koutsoudas, A., and Korfhage, R. "Mechanical Translation and the Problem of Multiple Meaning," Mech. Transl. Journal, Vol. III, No. 2, 1956.

2. Shannon, C. E., and Weaver, W. The Mathematical Theory of Communication, University of Illinois Press, 1949.

3. Zipf, G. K., Human Behavior and the Principle of Least Effort, Addison-Wesley Press, Inc., 1949.

4. Joos, M. Lang. 12, 1936, pp. 196-210.

5. Josselson, Harry H., The Russian Word Count, Wayne University Press, 1953.

6. Hanley, Miles L., Word Index to James Joyce's Ulysses, Madison, Wisc., 1937.

7. Shannon, C. E., "Prediction and Entropy of Printed English," Bell Syst. Tech. J., Vol. 30, pp. 50-64, 1951.

DISTRIBUTION LIST

(Unclassified)

Copy No.	
1-4	Office of Assistant Secretary of Defense (R and D), Tech. Library Branch, Department of Defense, Washington 25, D. C.
5	Special Assistant, (R and D), Office of the Under Secretary of the Army, Department of the Army, Washington 25, D. C.
6	Director, Weapons System Evaluation Group, Department of Defense, Washington 25, D. C.
7	Office, Assistant Chief of Staff, Intelligence, Training Division, Attn: Lt. Col. Paul E. Doherty, Department of the Army, Washington 25, D. C.
10	Chief, Army Security Agency, GAS-24, RD and TE, Attn: ASA Liaison Officer, Department of the Army, Washington 25, D. C.
11	Chief, Army Security Agency, GAS-15, Attn: ASA Liaison Officer, Department of the Army, Washington 25, D. C.
12	Commanding General, U. S. Army Combat Surveillance Agency, Department of the Army, Washington 25, D. C.
13	Office, Deputy Chief of Staff for Military Operations, Attn: Director, Plans Directorate, Department of the Army, Washington 25, D. C.
14	Office, Deputy Chief of Staff for Military Operations, O and T Directorate, Doctrines and Combat Developments Division, Department of the Army, Washington 25, D. C.
15	Office, Deputy Chief of Staff for Military Operations, O and T Directorate, Attn: Major Thomas R. Dolezal, Department of the Army, Washington 25, D. C.
16-17	Commanding Officer, Army Liaison Group (9550), Project MICHIGAN, Willow Run Laboratories, Ypsilanti, Michigan FOR TRANSMISSION TO Commander, British Joint Services Mission, 1800 K Street, N. W., Washington, D. C. THROUGH Office of the Chief Signal Officer, Attn: SIGRD-5-a, Department of the Army, Washington 25, D. C.
18	Chief, Tactics Division, Support Weapons Group, Operations Research Office, The Johns Hopkins University, 7100 Connecticut Ave., Chevy Chase 15, Maryland
19	Chief, Home Defense Division, Operations Research Office, The Johns Hopkins University, 7100 Connecticut Ave., Chevy Chase 15, Maryland
20	Lt. Col. A. W. Harris, USMC, Office, Assistant Chief of Bureau for Electronics (Code 803), Bureau of Ships, Department of the Navy, Washington 25, D. C.
21	Chief of Transportation, Attn: AVD-ED, Department of the Army, Washington 25, D. C.
22	Office, Chief of Research and Development, Attn: Lt. Col. W. M. VanHarlingen, Department of the Army, Washington 25, D. C.
23	Assistant Chief of Staff, G2, O and T Division, Attn: Lt. Col. R. V. Fridrich, U. S. Marine Corps Headquarters, Arlington Annex, Columbia Pike and Arlington Ridge Road, Arlington, Virginia
24	Chief, Aerial Reconnaissance Laboratory, Attn: Col. J. R. Knight, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio
25-26	Commanding General, Redstone Arsenal, Huntsville, Alabama

27	Commanding General, Frankford Arsenal, Attn: Chief, Research and Development Department, Philadelphia 37, Pennsylvania
28	Commanding General, Detroit Arsenal, Attn: Chief, Research and Development Division, Center Line, Michigan
29	Commanding General, Aberdeen Proving Ground, Maryland
30-35	Office, Chief of Ordnance, Research and Development Division, Department of the Army, Washington 25, D. C.
36	Commanding General, Transportation Research and Development Station, Fort Eustis, Virginia
37-38	Chief of Engineers, Attn: ENGNF, Department of the Army, Washington 25, D. C.
39	Chief of Engineers, Attn: ENGIE, Department of the Army, Washington 25, D. C.
40	Chief of Engineers, Attn: ENGIS, Department of the Army, Washington 25, D. C.
41-42	Director, Engineer Research and Development Laboratories, Attn: Chief, Electrical Engineering Department, Fort Belvoir, Virginia
43	Director, Engineer Research and Development Laboratories, Attn: Chief, Topographic Engineering Department, Fort Belvoir, Virginia
44	Director, Engineer Research and Development Laboratories, Attn: Chief, Military Engineering Department, Fort Belvoir, Virginia
45	Commandant, The Engineer School, Fort Belvoir, Virginia
46	Office of the Quartermaster General, Building T-A, 2nd and T Streets, S. W., Department of the Army, Washington 25, D. C.
47-55	Chief, Research and Development Division, Office of the Chief Signal Officer, Department of the Army, Washington 25, D. C.
56	Commander, Chemical Corps Research and Development Command, Department of the Army, Washington 25, D. C.
57	W. J. Merchant, Chief Int. Div., Operations Research Office, The Johns Hopkins University, 7100 Connecticut Ave., Chevy Chase 15, Md.
58	Commanding General, Signal Corps Engineering Laboratory, Attn: Mr. H. P. Hutchinson, Fort Monmouth, New Jersey
59	Commanding Officer, Army Map Service, Attn: Documents Library, 6500 Brooks Lane, Washington 25, D. C.
60	Commanding General, Army Electronic Proving Ground, Attn: Lt. Col. S. H. Webster, Battle Area Surveillance Department, Fort Huachuca, Arizona
61	Commanding Officer, Human Research Unit Nr. 1, Attn: Dr. W. D. Boiers, Fort Knox, Kentucky
62	Commanding General, Army Ballistic Missile Agency, Attn: ORDAB-T, Huntsville, Alabama
63	Directors, Human Resources Res. Office, Attn: Security Officer, George Washington University, P. O. Box 3596, Washington 7, D. C.
64	Fred Thompson, c/o Combat Development Department, General Analysis Corporation, Fort Huachuca, Arizona
65	General Analysis Corporation, Attn: Alex Mood, Santa Monica, California

72-96	Commanding General, Signal Corps Engineering Laboratory, Attn: SIGEL-DR, Fort Monmouth, New Jersey
97	President, Signal Corps Board, Fort Monmouth, New Jersey
98	Commander, Air Force Armament Center, Attn: Mr. Henry Maulshagen, Eglin Air Force Base, Florida
99	Haller, Raymond and Brown, Attn: SCEAG, University Park, Pennsylvania
100	Commanding Officer, Signal Corps Electronics Research Unit, (9560), P. O. Box 205, Mountain View, California
101-102	Commanding General, Army Electronic Proving Ground, Attn: Chief, Battle Area Surveillance Department, Fort Huachuca, Arizona
103	Dr. W. R. G. Baker, Vice President and General Manager, Electronics Department, General Electric Company, Electronics Park, Syracuse, N. Y.
104	Aerojet-General Corporation, Attn: Librarian, P. O. Box 296, Azusa, California VIA Bureau of Aeronautics Representative, c/o Aerojet-General Corporation, 6352 N. Irwindale Avenue, Azusa, California
106	Director, Research Studies Institute, Air University Command, Attn: Chief, ADTIC Maxwell Air Force Base, Alabama
107	Dr. Michael Ference, Jr., Ford Motor Company, P. O. Box 2053, Dearborn, Michigan
110	Dr. Albert G. Hill, Room 2E990, The Pentagon, Washington 25, D. C.
112	Dr. Andrew Longacre, Control Systems Laboratory, University of Illinois, Urbana, Illinois
117	Director, National Bureau of Standards, Attn: Division 12, Washington 25, D. C.
118-128	Commanding General, Continental Army Command, Fort Monroe, Virginia
129	Commanding General, Attn: CDTEC, Fort Ord, California
133	Commanding General, AA and GMC, Fort Bliss, Texas
134	Commanding General, The Armored Center, Fort Knox, Kentucky
135	President, Continental Army Command Board No. 6, Fort Rucker, Alabama
136	Commanding General, The Infantry Center, Fort Benning, Georgia
137	Assistant Commandant, The Artillery School, AA and GM Branch, Fort Bliss, Texas
138	Commandant, The Armored School, Fort Knox, Kentucky THROUGH Commanding General, The Armored Center, Fort Knox, Kentucky
139-140	Assistant Commandant, The Artillery and Guided Missile School, Fort Sill, Oklahoma
141	Commandant, The Infantry School, Fort Benning, Georgia
142-144	Commandant, Army War College, Carlisle Barracks, Pennsylvania
145-147	Commandant, Command and General Staff College, Fort Leavenworth, Kansas
148-149	President, Army Intelligence Board, Army Intelligence Center, Fort Holabird, Maryland
150	Commandant, Army Aviation School, Fort Rucker, Alabama

151	Commandant, Joint Air Support School, c/o Commanding General, Continental Army Command, Fort Monroe, Virginia
152-153	President, Continental Army Command Board No. 1, Fort Sill, Oklahoma
154	President, Continental Army Command Board No. 2, Fort Knox, Kentucky
155	President, Continental Army Command Board No. 3, Fort Benning, Georgia
156	President, Continental Army Command Board No. 4, Fort Bliss, Texas
157	President, Continental Army Command Board No. 5, Fort Bragg, North Carolina
158	Army Section, Marine Corps Equipment Board, Quantico, Virginia
159	Commanding General, Continental Army Command, Attn: Col. William M. Slayden, GS (Armor), Fort Monroe, Virginia
160	Chief, Office of Naval Research, Building T-3, Constitution Avenue between 16th and 17th Streets, Department of the Navy, Washington 25, D. C.
161-162	Director, Naval Research Laboratories, 4th and Chesapeake Streets, S. W., Department of the Navy, Washington 25, D. C.
163	Chief of Naval Operations (OP-37), Department of the Navy, Washington 25, D. C.
164-168	Chief, Office of Naval Research, Building T-3, Attn: Col. W. N. Flournoy, USMC, Department of the Navy, Washington 25, D. C.
169-170	Commandant of the Marine Corps, Arlington Annex, Columbia Pike and Arlington Ridge Road, Arlington, Virginia
171-172	Director, Marine Corps Development Center, Quantico, Virginia
173	Commander, United States Air Force Security Service, Attn: OPL, San Antonio, Texas
174-175	Department of the Air Force, Headquarters USAF, Attn: AFDRQ, Washington 25, D. C.
176-177	Department of the Air Force, Headquarters USAF, Attn: AFDRD, Washington 25, D. C.
178-179	Department of the Air Force, Headquarters USAF, Attn: AFDAP, Washington 25, D. C.
180-181	Department of the Air Force, Headquarters USAF, Attn: AFOAC, Washington 25, D. C.
182-183	Department of the Air Force, Headquarters USAF, Attn: AFOOP, Washington 25, D. C.
184-185	Department of the Air Force, Headquarters USAF, Attn: AFOIN, Washington 25, D. C.
186-200	Commander, Wright Air Development Center, Attn: WCLRO (Staff), Wright-Patterson Air Force Base, Ohio
201-202	Commander, Air Force Cambridge Research Center, Cambridge, Massachusetts
203-204	Commander, Rome Air Development Center, Air Research and Development Command, Attn: RCSST-3, Rome, New York
205	Commander, Air Force Armament Center, Attn: ACOTT, Eglin Air Force Base, Florida
206	Director, Research Studies Institute, Air University Command, Attn: Chief ADTIC, Maxwell Air Force Base, Alabama
207-211	Documents Service Center, Armed Services Technical Information Center, Knott Building, Dayton 2, Ohio
212	Chief, Armed Forces Special Weapons Project, Attn: Adjutant General, Room 1B662, Pentagon Building, Washington 25, D. C.
	Headquarters, USAF, ATTN: AFOIVI-3AIEI, Temop U Bldg., 12th and Constitution N. W., Washington 25, D. C.

218-219	Commander, Headquarters Ninth Air Force, Attn: DCS/O, Shaw Air Force Base, South Carolina
220-221	Commander, Air Technical Intelligence Center, Wright-Patterson Air Force Base, Attn: Mr. R. B. Keeney, ATIR, Dayton, Ohio
222-223	Commander, Air University, Maxwell Air Force Base, Montgomery, Alabama
224-225	Commander, Air Proving Ground Command, Attn: Deputy for Operations, Eglin Air Force Base, Florida
228-229	Commander, Strategic Air Command, Offutt Air Force Base, Omaha, Nebraska
230-231	Commander, Far East Air Forces, APO 925, c/o Postmaster, San Francisco, California
232-233	CINC, United States Air Forces Europe, Attn: Director of Intelligence, APO 633, c/o Postmaster, New York, New York
234-235	The RAND Corporation, Attn: Mr. M. Davies, 1700 Main Street, Santa Monica, California
236-238	Headquarters, Tactical Air Command, Attn: Col. James L. Rose, Langley Air Force Base, Virginia
239-240	Commanding General, Air Research and Development Command, P.O. Box 1395, Attn: Lt. Col. J. J. Pellegrini, Baltimore, Maryland
241	Cornell Aeronautical Laboratory, Inc., Attn: Librarian, 4455 Genesee Street, Buffalo 21, New York
	THROUGH
	Bureau of Aeronautics Representative (Contract No. AF 18(600)-2), 4455 Genesee Street, Buffalo 21, New York
242	Pacific Division, Bendix Aviation Corporation, Attn: Mr. W. C. Leitch, Chief Electronics Engineer, 11600 Sherman Way, North Hollywood, California
	THROUGH
	Inspector of Naval Materiel, 1206 Santee Street, Los Angeles 15, California
243	Raytheon Manufacturing Company, Government Contracts Division, Attn: Documents Section, Waltham 54, Massachusetts
	THROUGH
	Assistant Inspector of Naval Materiel, Foundry Avenue, Waltham 54, Massachusetts
244	The Glenn L. Martin Company, Baltimore 3, Maryland
	THROUGH
	Air Force Plant Representative Office, WRAMA, The Glenn L. Martin Company, Baltimore 3, Maryland
245	Columbia University, Electronics Research Laboratories, Attn: Miss Helen K. Cressman, Technical Editor, 632 W. 125th Street, New York 27, New York
	THROUGH
	Commander, Rome Air Development Center, Attn: RCSST-3, Griffiss Air Force Base, Rome, New York
251	Commanding General, Detroit Arsenal, Attn: ORDMX-ECCT-Astron Project Engineer, Center Line, Michigan

252	Chief, Los Angeles Ordnance District, U. S. Army, Attn: ORDEV, Ord. Res. Br., 55 South Grand Avenue, Pasadena 2, California
253	General Electric Company, Attn: Miss Sophia Rugare, Librarian, French Road, Utica, New York THROUGH Office of the Air Force Plant Representative, MAAMA, General Electric Company, Electronics Park, Syracuse, New York
254	Commanding Officer, Army Liaison Group (9550), Project MICHIGAN, Willow Run Laboratories, Ypsilanti, Michigan
256	Office, Continental Army Command Liaison Officer, Project MICHIGAN, Willow Run Laboratories, Ypsilanti, Michigan
257	Corps of Engineers Liaison Officer, Project MICHIGAN, Willow Run Laboratories, Ypsilanti Michigan
258	Air Force Development Field Representative, Project MICHIGAN, Willow Run Laboratories, Ypsilanti, Michigan
259	Commander, Wright Air Development Center, Attn: WCOL-9, Wright-Patterson Air Force Base, Dayton, Ohio
261-310	The University of Michigan, Internal Distribution

AD______________________ Accession No.______________________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED

AD______________________ Accession No.______________________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED

AD______________________ Accession No.______________________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED

AD______________________ Accession No.______________________

UNCLASSIFIED

The University of Michigan, Engineering Research Institute, Willow Run Laboratories, Willow Run Airport, Ypsilanti, Michigan

Koutsoudas, Andreas M. and Machol, Robert E., Frequency of Occurrence of Words – A Study of Zipf's Law, with Application to Mechanical Translation

Report No. 2144-147-T, April 1957, 17 pp., 2 illus., Project 2144 (Contract DA-36-039 SC-52654, DA Project NR-3-99-10-024, Sig C No. 102D), UNCLASSIFIED

Existing laws concerning the frequencies of words in language – specifically Zipf's and Joos' laws – are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising.

1. Mechanical Translation
2. Mathematics
3. Linguistics
4. Combat Surveillance
5. Contract No. DA-36-039 SC-52654

UNCLASSIFIED