[From: *Language* 33(4), 1957]

MECHANICAL TRANSLATION AND ZIPF'S LAW

ANDREAS KOUTSOUDAS

University of Michigan

[A problem which arises in the course of research on mechanical translation is the prediction of dictionary size. This article investigates the relation between empirical frequency laws and the function V(n)—the expected number of different words in an n-word sample of text. It is found that the probability-law proposed by Joos (1936) yields results which do not check well with experiments, and it is concluded that some modification of it is necessary for the purpose of vocabulary prediction.]

In 1956 the University of Michigan began a program of research to investigate [the possibility of mechanically translating a segment of language.[1] The ultimate goal 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- or post-editor. It will be necessary to provide (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 words. All four, of these requirements must be satisfied in such a way that they can be easily acted upon by an electronic computer.

The linguistic requirements involved in enabling an electronic computer to handle the identification, syntax, and multivocality of a language universe have been described in an earlier paper.[2] Here we shall concern ourselves with one requirement for the ultimate successful mechanical translation of languages, that of the dictionary and its size.

An immediate economic problem arises with regard to the compilation of a dictionary. Can a dictionary, be compiled, for instance, which is large enough to accurately represent the language universe in question and small enough to be contained in an electronic memory? Specifically, can it be safely stated that the number of different words needed to translate, for example, Russian astronomy texts into correct and unambiguous English is small enough so as not to exceed the capacity of an electronic computer for a given period of time? For our 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 dic-

[1] This study was conducted at the University of Michigan under the author's supervision, with funds provided by the Engineering Research Institute. The author is obliged to George Minty for his co-operation in the writing of this paper.

The present article is primarily mathematical, in the sense that the solutions given, though linguistically significant, were obtained by mathematics. All the mathematical derivations underlying the article have been published in University of Michigan Report 2144-147-T, and are available on request.

[2] Koutsoudas, Mechanical translation and the problem of multiple meaning, *MT journal* 3:2 (1956).

tionary to be constructed in the course of developing mechanical translation, somewhat different definitions of a word will be necessary, but the present as sumption is most useful for using machinery to make counts of words in large samples of text.

The problem of dictionary size may be viewed essentially as the same problem that linguists have faced when called upon to write an elementary text for the teaching of a foreign language or to describe any segment of a given living language or dialect. Linguists will find themselves asking, How many words appear in the segment? Which words occur only once or twice? Which are the most important? How often will new words be introduced into the segment? To answer such questions, the linguist has had to collect raw data: large and different samples of discourse. It would be advantageous to establish laws which state that a segment of language will behave in such and such a way in regard to the frequency of occurrence of a particular word. It is not surprising, therefore, that in order to predict the dictionary size needed for mechanical translation, we were forced to investigate several frequency studies made on large samples of text and consequently to examine, in theory at least, the validity of the various laws established from these frequency studies. The invention of tools to test their validity became a second problem for our consideration.

With respect to this problem we set out to develop a new mathematical method which correlates the frequency-vs.-rank relationship of words with a second relationship: number of words vs. number of different words. Having established this method, we can take empirically established frequency laws (for example those of Condon, Zipf, Yule, Mandelbrot) and, by treating them with this mathematical method, either confirm or refute their validity. We are claiming, consequently, that the linguist will be provided with a means not only of predicting a vocabulary but of checking extrapolated empirical laws by comparison of their mathematical consequences with easily obtainable experimental results.

After the consideration of these two problems, we are prepared to introduce the empirical laws and theoretical models of Zipf[3] and Joos.[4] In the process, a description will be offered of the tool used to investigate their consequences, namely, the 'probabilistic model'.

In studying human behavior as a natural phenomenon, Zipf devoted a great amount of effort to analyzing language statistically. The motivation behind this study was, of course, Zipf's strong desire to establish laws which could successfully describe complex areas of human behavior, of which language is the most important.

The statistical analysis performed on a large amount of data collected by Zipf resulted in at least one significant contribution: the establishment of a general law which could describe many aspects of human behavior.[5] With respect to language the law reads, 'The product of the relative frequency of occurrence of a word and its rank is equal to a constant *(F* X *R = C).'*

[3] *G.* K. Zipf, *Human behavior and the principle of least effort* (Cambridge, Mass., 1949).

[4] M. Joos, review of *Zipf's The psycho-biology of language* in *Lg.* 12.196-210 (1936).

[5] Zipf's basic premise is that there are certain underlying principles controlling such sociological factors as the use of words, and that these principles are similar over a wide variety of social phenomena.

Although this law was stated in many different ways, Zipf identified it with the mathematical harmonic series (= 1 + ½ + ⅓ + ¼ + ...); indeed, the flaw is commonly known as the 'harmonic series law'. Whatever qualifiers we may attach to it are of no important consequence as long as we understand the kind of information that it conveys. Let us then state Zipf's law in more precise terms. According to Zipf, it is an empirical fact that in a large sample of running words, the 'rank' of a word—that is, the ordinal number of the word when all the different words of the sample are arranged in decreasing order of their frequency —is inversely proportional to its relative frequency—that is, the number of times the word occurs divided by the total number of words in the sample. In order that the inverse proportionality can be written as an equation, a constant of proportionality is needed. Zipf noticed that the products $F \times R$ were all on the order of 1/10. The law then takes the following symbolic form:

$$F = 1/10R \tag{1}$$

It should be noted that the sum of the relative frequencies of all the different words in the sample must equal 1, from the way relative frequency was defined. We cannot naively assume that the sum of the approximate relative frequencies 1/10R over all the words in the sample is exactly 1, otherwise we could conclude that the sample contains 12,000 different words regardless of its length.

We have only to cite now under what conditions the law holds and thus complete our explanation of it. Zipf's law is subject to the following three restrictions: (1) the sample of words must be large, (2) the law does not hold for the high-frequency words, and (3) the law does not hold for the low-frequency words. Strictly speaking, then, Zipf's law holds only for words of intermediate frequency of occurrence in a large sample.

Since our study is actually based more directly on Joos's 'probabilistic model',[6] we now consider Joos's interpretation of Zipf's Law.

In Joos's model, the words of a message are chosen independently, each by the same random process. The random process can be characterized by a decreasing probability function *P(R),* where *R* now stands for the rank of the word IN AN AUTHOR'S VOCABULARY (not in a particular sample of his work) and is defined as the ordinal number of the word when the words of his vocabulary are arranged in order of decreasing probability of use. Note that the sum of all probabilities must be 1; i.e., if *M* is the total number of words in the author's vocabulary,

$$P(1) + P(2) + \ldots + P(\mathrm{M}) = 1. \tag{2}$$

This idea is exemplified by a Wheel of Fortune, a circle whose total area is 1, and which is divided into segments, each labeled with a word. To decide on the first word of a text, the 'author' spins the arrow and writes down the word to which it points; for the second word of the text, he repeats the process; and so on. For convenience, we arrange these wedge-shaped areas around the circle in order of decreasing size, and label them with numbers in this order. These labels

[8] It should be noted that this model was used explicitly by Shannon and Weaver in their development of communication theory, and only implicitly by Joos (1936) in his interpretation of Zipf's Law.

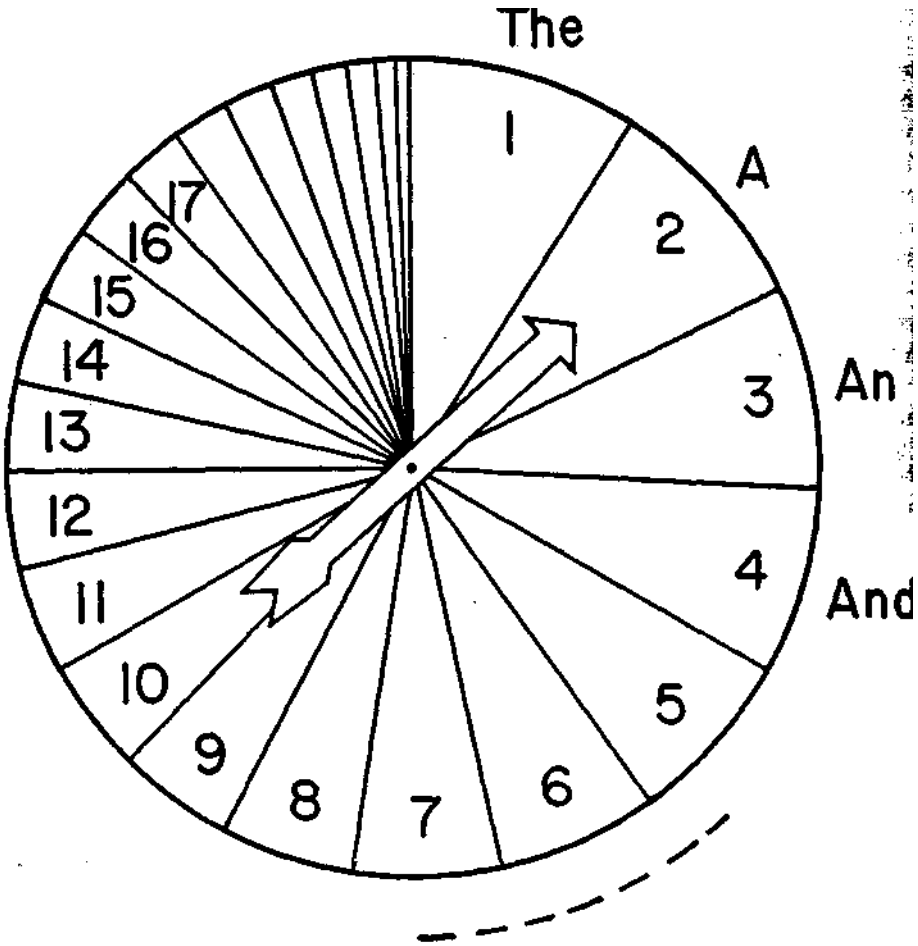

FIGURE 1

are referred to by the letter *R.* The biggest area is called *P*(l); the next biggest *P(2),* and so on. *P(R)* is the area which is numbered *R;* that is, *P(R)* is the area of any wedge-shaped area around the circle.

Joos correlates his probabilistic model with Zipf's Law as follows: he observes that Zipf's empirical law is precisely what one would expect to find if *P(R)* is approximately 1/10R (except for the few most frequently used words). But he objects to Zipf's Law on empirical grounds and to this approximation for *P(R)* on theoretical grounds. First, he notes that Zipf's data are-better fitted by the approximation

$$F(R) = 1/10R^{\alpha} \tag{3}$$

where α is a constant slightly greater than 1. Second, he remarks that since it takes about 12,000 terms of the series 1/10 + 1/20 + 1/30 + … to equal 1,1/10*R* can be a probability function only if the author's vocabulary is about 12,000 words—a very unnatural restriction.

Accordingly, he proposes formula (3), which we shall call Joos's empirical law, to replace Zipf's Law, and also what we shall call Joos's assumption:

$$(4 \text{ and } 5) \qquad P(R) = \frac{1}{10R^{\alpha}}, \quad \sum_{R=1}^{M} \frac{1}{10R^{\alpha}} = 1$$

to explain his empirical law on the basis of this probabilistic model.[7] He does not require M to be 12,000, but does require that whatever value is chosen for M, the value of a must be chosen accordingly to satisfy formula (5).

We can offer an additional explanation by referring to the Wheel of Fortune of Figure 1 and asking how we are going to divide up this wheel, or rather, how we are going to determine the size of each wedge-shaped area around the circle. If we try to use Zipf's Law $P(R) = 1/10R$, so that the areas are 1/10,1/20,1/30, and so on, we will cover the wheel with about 12,000 words. Joos suggests that we take $P(R) = 1/10R^{\alpha}$ in order to obviate this peculiar restriction. By having α as exponent of R, we can make the wedge-shaped areas smaller and thereby increase their number to more than 12,000, or make them larger and decrease their number to less than 12,000. For example, if α is greater than 1.106, we will not cover the whole wheel area even with an infinite number of words. If α is exactly 1.106, it will take an infinite number of words to cover the area. If α is less than 1.106, we will cover the area with a finite number of words. If $\alpha = 1$, we can do it with 12,000 words (Zipf's Law). From these examples we can observe that α is related to the total vocabulary, which we have called M. We can choose α and then determine M from it,-or alternatively choose M and determine α from it.

Our work is based on Joos's probabilistic model, with Joos's assumption as a tentative hypothesis.

A formula for the expected value, which we shall call $V(n)$, of the number of different words in an n-word sample is easily derived in Joos's probabilistic model. Consider one of the words of the n-word message. The probability that it is the word of rank R is $P(R)$; the probability that it is NOT the word of rank R is $1 - P(R)$. The probability that none of the words of the message is the word of rank R is $[1 - P(R)]^n$. The probability that at least one of the words of the message is the word of rank R is therefore $1 - [1 - P(R)]^n$. The expected value of the number of different words in an n-word sample is therefore, according to elementary probability theory,

$$(6) \qquad V(n) = \sum_{R=1}^{M} 1 - [1 - P(R)]^n,$$

If we now hypothesize, with Joos, formulas (4) and (5), we obtain the pair of formulas (5) and (7):

$$(7) \qquad V(n) = \sum_{R=1}^{M} 1 - \left(1 - \frac{1}{10R^{\alpha}}\right)^n$$

[7] Sigma is a mathematical symbol denoting a summation process. It is read, 'the summation of'.

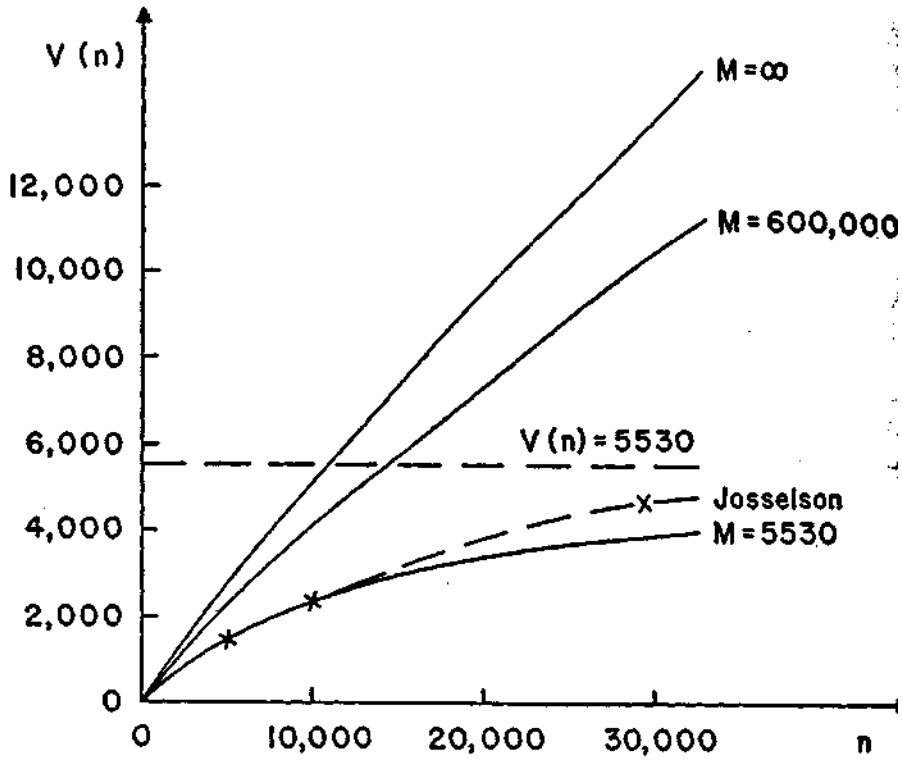

FIGURE 2

$$1 = \sum_{R=1}^{M} \frac{1}{10R^{\alpha}} \tag{5}$$

We used these formulas to attempt to predict the results of word counts made by Josselson on 5,000-word, 10,000-word, and 29,345-word samples from Pushkin's *The Captain's Daughter*[8] Before the formulas could be applied, we had to make some assumption as to the value of the constant *M*. The simplest and most natural assumption is $M = \infty$. Another natural trial-assumption is $M = 600{,}000$ (the approximate number of words in a dictionary the size of *Webster's new international dictionary).* From formula (6), $\alpha = 1.106$ and $\alpha = 1.1$ respectively for these values of *M*. Graphs of *V(n)* obtained from formula (7) are shown in Figure 2. As this figure shows, the theoretical predictions disagree markedly with the actual word counts.

The next approach we tried was to choose *M* so that the graph would have to pass through the experimental point for $n = 10{,}000$. The choice of *M* here is $M = 5530$. The graph of *V(n)* obtained is shown in Figure 3. It is asymptotic to (i.e. approaches but never reaches) the line $V(n) = 5530$. Although this theoretical curve fits the experimental values much more closely than the first two, it still fails for one of the points. It would not fail if Joos' assumption was

[8] H. Josselson, *The Russian word count* (Detroit, 1953).

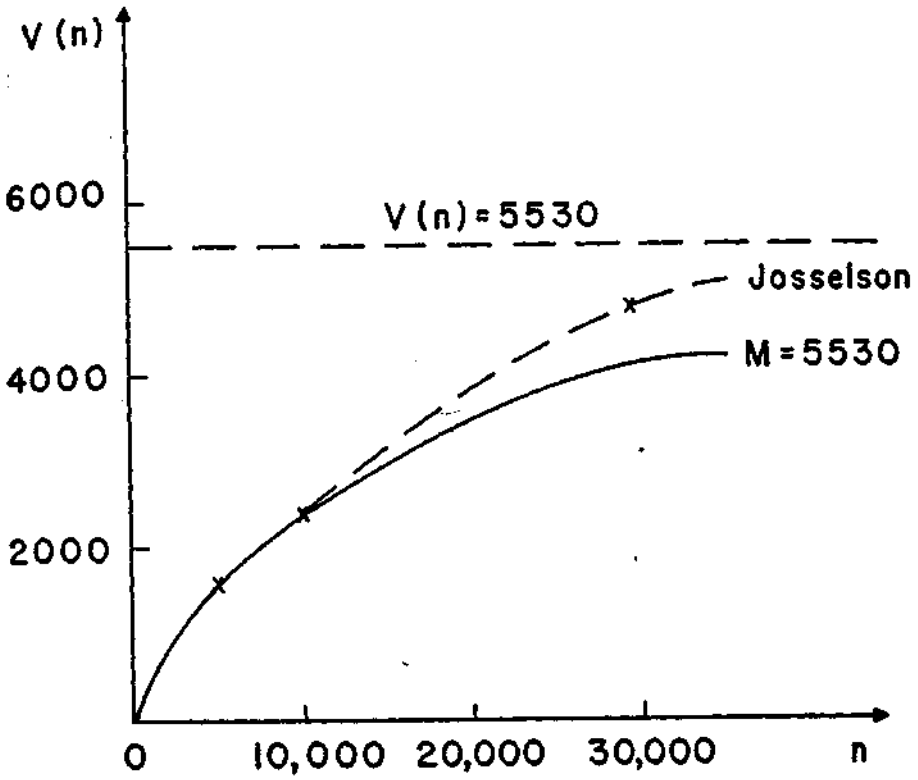

FIGURE 3

correct. Apparently the basic reason for this disagreement is that the theoretical graph never goes above $V(n) = 5530$, while it seems intuitively certain that if a bigger sample were taken, Pushkin's vocabulary would be found to be much larger than 5530 words.

In re-examining Joos's assumption to find possible reasons for the disagreement with experiment, we came to the conclusion that the restriction (5) is fundamentally a fallacy. For the sum of the probabilities given by the approximation-formula is not equal to 1, but is equal to 1 plus the sum of the errors of approximation. We should not require this latter sum to be zero, since we do not require the approximation to be good for small R. In other words, the restriction (5) introduces much too strong a coupling between M and the errors in the approximation-formula for small R. Thus M and α should be regarded as unrelated, and values for them should be decided separately.

We feel, too, that in the assumption $P(R) = 1/aR^{\alpha}$, the constant 'α' should be determined as exactly as possible, not chosen as 10 for simplicity's sake alone.

We have not made any attempt as yet to see if the dropping of the requirement (5) results in an improved agreement of theory with experiment. We hope, however, to test this idea against a Russian word count now being conducted at the University of Michigan.

The author's general conclusions are as follows: Joos's probabilistic model and the formulas derived from it show very considerable promise as a means of correlating empirical frequency laws with word counts, and also, taken together with an adequate probability function, as a means of predicting vocabulary for the purpose of mechanical translation. Joos's assumption is inadequate for this purpose, but perhaps some modification of it can be found which will be satisfactory.