SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Statistical Macro-Linguistics.

B. MANDELBROT

Institut de Mathématiques - Université de Lille

It is well-known, how a simple and economic theory may transform an empirical law from something quite amazing and difficult to believe, into something almost obvious and even trivial. It seems that such will be the final fate of certain laws of linguistics; the relationship between rank and frequency for natural words, and the relationship between species and genera in natural taxonomies. Let us recall that these laws were discovered by J. B. ESTOUP and J. C. WILLIS, respectively, but were made well known by the publications of G. K. ZIPF [1]. The author's models for these results were published since 1951, and their final form was given in 1957. Although these theories are essentially very simple, we have not yet found a way of developing them fully in a few pages. We shall therefore limit ourselves in this Note to a bare outline of the theory of the frequency distribution for natural words.

The first main tool of the theory is the following relationship:

$$C = -\log_2 p \,,$$

where p is the probability of occurrence of some signal in a message and C is the « cost » of transmitting this signal in some optimal binary code. This relationship is extremely familiar in information theory and may be obtained under a wide variety of definitions of optimality; we shall not attempt here to reduce this relationship to more fundamental concepts. Further, we shall not restrict ourselves to binary codes, and shall write:

$$\beta C = -\log_e p \,, \tag{1}$$

where β is a factor which depends upon the scale chosen for C.

Let us apply the relationship (1) to the words of natural language. Each word will be labelled by the rank, which it occupies in a list of all words, arranged by order of decreasing probability in a given text: that is, $r=1$ designates the most frequent word, $r=2$, the second most frequent, etc.; the number of words more frequent than a word of frequency p will be $r(p)-1$. Then, the empirical result is that for words other than the most frequent ones (large r) one has, whichever the language in which a test was written:

$$p(r) = Pr^{-B}, \tag{2}$$

where P and B are some constants. The relationship (1) then becomes:

$$\beta C = -\log P + B \log r,$$

$$\log r = \frac{\log P}{B} + \frac{\beta}{B} C = \log K + \beta' C,$$

(by definition of K and β'); finally,

$$r = K \exp[\beta' C], \tag{3}$$

An « explanation » of the law of Zipf requires an interpretation of the « cost » of coding a word and a model for the structure of the word, which together would lead to (3). One reasonable interpretation of « cost » would be the number of letters required for the code. It turns out actually that this interpretation cannot be carried to the end, and one must rather think of the cost as being something like the time required to read a word [2]. However we shall sketch a theory based upon the identification of cost to (essentially) the number of letters. The second step is the choice of the rule of formation of words: in the present model, one will assume that a word is any sequence of letters contained between successive occurrences of some additional improper letter, the « space ».

It is then reasonable to interpret cost as being equal to the number of proper letters, plus the cost C_0 of the improper letter « space ». Let there be M different proper letters. Then

there is	1	word	of cost C_0
there are	M	words	of cost C_0+1
there are	M^2	words	of cost C_0+2, etc.

Adding, one finds that

$$\text{there are } \frac{M^n-1}{M-1} \text{ words of cost less than } C = C_0 + n.$$

For large n, this gives

$$r = K' M^{C-C_0} = K \exp [C \log M]$$

which is of the form required to explain Zipf's data on word frequency for large r.

It is unfortunate that the simplest case above cannot be carried out to further steps without some difficulties. However, it turns out that the same result (3) can be obtained under wider, more realistic and mathematically more convenient conditions, as long as *a word is a sequence of letters contained between two successive spaces, as long as there is « little interaction » between successive proper letters, and as soon as one can justify* (1).

A closer examination of the cost of coding for small values of r suggests the following improvement of the law (2), valid for all r,

$$p(r) = (B - 1) V^{B-1} (r + V)^{-B},$$

where V is a second coefficient. This further approximation turns out to be experimentally excellent (*).

(*) Other explanations may eventually be given of the rank frequency relation (2). However, the explanation suggested by H. A. Simon is certainly incorrect. See our *Note on a class of skew distribution functions*, in *Information and Control*, **2**, 90 (1959).

REFERENCES

[1] G. K. Zipf: *Human Behavior and the Principle of Least Effort* (Cambridge, Mass. 1949).

[2] B. Mandelbrot: *Linguistique statistique macroscopique*, I. One of the three essays in the book *Logique, langage et théorie de l'information*, by L. Apostel, B. Mandelbrot and A. Morf (Paris, 1957), and B. Mandelbrot: *Linguistique statistique macroscopique*, II. A report of the Institut de Statistique de l'Université de Paris (Paris, 1957).