ON RECURRENT NOISE LIMITING CODING*

Benoit Mandelbrot**
Laboratoires d'Electronique et de Physique Appliquées,
Paris, France

Summary.

A method of word by word coding can be described by a coding tree. The study of the coding methods is equivalent to the study of their trees, considered as graphs. The number of letter strings used as codes (spellings), considered as a function of the length C of these strings, is by definition the "structure function" S(C) of the tree and of the coding method. Two coding methods having identical structure functions lead to the same cost of coding for any message, and are called equivalent. - A coding method is said to be recurrent if the decision as to whether a letter is a last letter of a word requires the knowledge of the preceding letters of this word only. The structure function of a recurrent method of coding satisfies Szilard's inequality: $\Sigma S(C)\, M^{-C} \leqslant 1$, where M is the number of letters and to any function satisfying this inequality corresponds at least one recurrent coding method. Sardinas and Patterson have shown that there are cases in which the message may be recovered from the coded string of letters, even though the identification of the ends of words requires the knowledge of the future of the message. However, the structure functions of these coding methods must still satisfy Szilard's inequality, and they can always be replaced by an equivalent recurrent coding procedure. No advantage is to be gained from non-recurrent codes using the future.

In addition to Shannon's problems of coding without noise or in the presence of finite noise, one introduces the problem of coding in the presence of a very small noise. One then imposes the requirement that an error occurring in one letter does not destroy the whole message. This requires either that all coded words be of same length, or that they end up with letters which do not occur inside of the word.

If there is equality in Szilard's relation there exists one language which can be coded by an uncorrelated letter sequence. It is said to be matched to the coding method, which is said to be complete. In particular, complete error-limiting codes have as structure functions,

* Presented at the SYMPOSIUM ON INFORMATION NETWORKS, Polytechnic Institute of Brooklyn, April 12-14, 1954.
The research on which this paper is based was supported in part by the Rockefeller Foundation.

** Presently at the Institute for Advanced Study, Princeton, N. J.

respectively, M^{C_o} for one value of C_o, and 0 elsewhere; and $(M-M')\,M'^{C-1}$, where $M' < M$. The languages matched to these structure functions are, respectively, composed of equiprobable words; and of words following, except for small ranks, the law $p_r = P\,r^{-B}$, where r is the rank by decreasing frequencies, and $B = (\log M)/(\log M')$.

In the general case, there is no tree built with an arbitrary given alphabet, to which an arbitrary given language is matched in an error limiting fashion. Let us, however, generate new words for the language by taking either fixed numbers, given a priori, of the old words; or all the sequences of the old words contained between two successive returns to any fixed old word, chosen to be the "space ". Shannon's theory of coding, and the definition of the information, are based upon the fact that the new words following the first definition are essentially equiprobable (law of large numbers), quite irrespective of the generating words, and of their independent or Markoff character. The only function of the generating words is the information logarithm of the number of generated words.

If one takes instead the second method of generating words, one can prove a sort of new "recurrent" limit theorem, which shows that the words under the new definition are again independent of the generating words, except for a single number B. That is, they follow the above law $p_r = P\,r^{-B}$. Information does not appear and is not needed; its role is played by the ratio $1/(B-1)$. This new limit theory may be used, instead of the conventional one, to construct an alternative theory of coding, fully equivalent to Shannon's as far as the theorems are concerned, but more appropriate for some applications. The bulk of the paper will be given to this theory.

In particular, it seems to be established that the words in natural languages, in the usual definition of a word as anything between two spaces, do follow the limit statistics corresponding to the generation of new words from any initial word system, by our second method. Therefore, natural languages as sequences of words do transmit as much information per word as is compatible with recurrent coding with space and a limitation in the average cost of coding per word.

1. On Recurrence In Coding.

1.1 Definitions. A discrete finite stationary irreducible (ergodic) one-dimensional random process may be considered as a statistical language, and any sequence of realizations is then called a message. The elements of the message will be called the words W_r $(1 \leqslant r \leqslant R)$, the set of words is the vocabulary $\mathcal{W}$, and the language will also be called the W-process. A statistical language need not be a natural language, but it may very well be considered as a model of one. We have found in previous investigations (1953)(1954a) that the theory of statistical languages is very useful in the study of natural languages; these results will be incorporated in a generalized and summarized form in

the present paper. Reciprocally, the structure of natural languages may be of great inspiration in the study of more general languages: in particular, we shall see the role which can be played by the concept of "space".

A word by word code for a message is a string of other elements, called the letters L_g, taken out of an alphabet $\mathcal{L}$, such that any word in a given position or context is given a well-defined spelling. The sequence of letters will be called the L-process. It is uniquely determined by the W-process and a method of coding. We assume that there is no noise.

A coding procedure is fully described by its coding tree. This is a rooted tree, in which the degree (or ramification number: the number of lines ending up at the vertex) of each vertex is 1 or M+1, where M is the number of letters, except for the distinguished root, which has degree M. Some, but not necessarily all, of these vertices represent word codes. Vertices of degree 1 are terminal vertices. The discussion of different coding procedures is in fact equivalent to the discussion of various properties of the coding trees considered as graphs, and in particular of the distribution of the vertices distinguished by the fact that they correspond to admissible codes for words. In all cases, when the number of vertices representing word codes is considered as a function of the distance from these vertices to the root, one obtains by definition the structure function S(C) of the coding method and of its tree. This term is borrowed from Khinchin (1949), whose exposition of statistical mechanics seems the best frame of reference for pointing out the formal identity between our later considerations and those of statistical mechanics. Two coding procedures having equal structure functions lead to the same cost of coding for every message, and may be called equivalent, even when their coding trees are not isomorphic. The cumulated sum from 1 to C of the structure function is the rank function r(C). Sometimes, the admissible coding letter strings will be ranked by increasing length and designated by C_r.

1.2. Recurrence; Szilard's Relation and Exponential Average Relation*.

One must be able to recover the W-process from the L-process alone, without outside help. Within this condition, one wants to choose the method of coding leading to the least cost of coding in average number of letters per words. (Later, cost will be defined in more general terms.) The original methods for the matching of a coding procedure to the language statistics were proposed by Shannon (1948) and Fano (1949), and their definitive improvement is due to Huffman (1953). In all these methods, one uses the condition that, all the preceding words being known, the spelling of a word is never the beginning of the spell-

* The results of Sec. 1.2 are not new, except for details of presentation, and they are repeated only as an introduction to what follows.

ing of another word. Suppose that, moreover, the spelling does not depend upon the preceding words, that is, that no account is taken of the joint probabilities of the words in the matching of the coding procedure. Then the event in the L-process, expressed by the fact that a letter is a last letter of a word, is determined by the letters having occurred since the end of the preceding word. No knowledge of the forward or of the backward context is necessary to see whether a word is ending. We propose to call recurrent all methods of coding of the above type which do not require the knowledge of the context to decide whether a sequence of letters starting at the end of a word is itself a word. This definition is slightly different from that of Feller (1950), in which one should be able to decide whether a sequence of letters ends at the end of some word even if the sequence does not start at the beginning of a word. However, our definition of recurrence will coincide with Feller's in the most important case, our second error-limiting case.

If a method of coding is recurrent, all vertices of the coding tree which represent words are terminal vertices, or vertices of degree 1. It follows that the structure function of such a tree must satisfy the inequality

$$\sum_{c=1}^{\infty} S(C)\, M^{-C} \leqslant 1$$

This inequality was introduced by Szilard (1929) (in a structurally identical problem relating to Maxwell demons), and rediscovered many times (in particular by L. K. Kraft in an unpublished M. I. T. thesis, 1949). To prove the inequality, consider a number L, and complete the coding tree so that no vertex is at a distance from the root smaller than L. Each of the S(C) vertices at a distance C from the root gives M^{L-C} vertices at the distance L. As the total number of vertices at distance L cannot exceed M^L, one has, for every L, $\Sigma_1^L S(C)\, M^{L-C} \leqslant M^L$. Dividing by M^L, one gets $\Sigma_1^L S(C)\, M^{-C} \leqslant 1$. The inequality remains true when L tends to infinity, q. e. d.

Reciprocally, it is sufficient for the existence of at least one coding tree of structure function S(C) that S(C) satisfy Szilard's inequality. To prove this, proceed to construct such a coding method step by step. Let C_o be the smallest C for which $S(C) \neq 0$. Consider the tree having all its M^{C_o} terminal vertices at the distance C_o from the root. Select any $S(C_o)$ of these vertices; this is possible since $S(C_o) \leqslant M^{C_o}$. Continue all other vertices down to distance $C_o + 1$ from the root. Select $S(C_o + 1)$ of these vertices; it is possible since their number is $M(M^{C_o} - S(C_o)) \geqslant S(C_o + 1)$. One continues by induction.

Szilard's inequality may be re-expressed in two ways:

One may define F by $\Sigma S(C)\, M^{-C} = M^{-F}$. Then one must have $F \geqslant 0$. This F is a sort of counterpart of the "free energy" of thermodynamics.

One may also define the number E such that $\sum_r R^{-C_r/E} = 1$ where R is the number of numbers C, supposed ranked in an arbitrary order. This number E is unique by Descartes' rule on the number of positive roots of an equation. We propose to call it the Exponential average of the numbers C_r. If all the C_r are equal, their average is equal to their common value, as it should be. The definition is of course inspired by those of the arithmetic average $\sum_r C_r/E = 1$, of geometric average $\prod_r (C_r/E) = 1$, of harmonic average $\sum_r E/C_r = 1$. With this definition, the Szilard's inequality becomes the "exponential average" inequality

$$\text{Exp, Av. (lengths of codes)} \geqslant \log_M R.$$

which is particularly perspicuous.

If, reciprocally, all terminal vertices of a coding tree represent admissable codes, the corresponding coding method and coding tree will be called complete. The Szilard and the exponential average relations are then equalities. Binary Fano and Huffman codes are complete, Huffman codes of basis M are complete if R-1 is a multiple of M-1, and Shannon codes are not necessarily complete. Complete codes are characterized by the fact that there exists one language, called reciprocally matched to or conjugate to the coding method, which, when coded by this method, gives an uncorrelated sequence of letters. In this language, all probabilities are of the form M^{-C}, and the number of words of probability M^{-C} is S(C).

If a language is not matched to any code of basis M, that is, if the $\log_M p$ are not all integers, one still sees that the set of numbers $[-\log_M p]$, equal to the smallest integers greater than $-\log_M p$, still follows Szilard's inequality. Shannon's method of coding gives explicitly a method of coding equivalent to the $[-\log_M p]$. It follows that the cost of coding of any message should never exceed the value it takes for the system $[-\log_M p]$. This value is $\sum p_r [-\log_M p_r] \leqslant -\sum p_r \log_M p_r + 1$. Let now a word system code C_r be incomplete, that is, let $F > 0$. For any given language, when coded with this method of coding, $\sum p_r C_r \geqslant -\sum p_r \log p_r + F$. Therefore, if F is also > 1, the method of coding C_r can be immediately improved upon by replacing C_r by $[-\log_M P_r]$. In other terms, to Szilard's inequality, which is a logical necessity, one may add the inequality $F < 1$, or

$$\text{Exp. Av. (word lengths - 1)} < \log_M R,$$

which is a pragmatic restriction.

We have thus introduced the expression $H = -\sum p_r \log_M p_r$, which is such that the cost of coding is always between H and H+1. H is of course the information, but it was important to introduce it, not through the consideration of all possible permutations of the words as Shannon does, but by the requirement of recurrence, which makes it impossible to take account of anything but the absolute probabilities of

the words while minimizing the cost of coding. This way of introducing information is important later on in the utilization of this concept in language theory. This is why we insisted on proving Szilard's inequality (as well as the equivalent inequality on the exponential average) without using the concept of information.

1.3. Non-recurrent Coding. Two questions are to be asked about them: are they conceivable, and if so, are they useful, that is, can they decrease the cost of coding word by word? Non-recurrence backward and forward must be considered separately.

Shannon has pointed out that non-recurrence backward, that is, utilization of the previous words in the delimitation of the current word, can be used to decrease the cost of coding Markoff messages. The spelling is chosen, not on the basis of the absolute probabilities but of the probabilities given the known preceding words. Therefore, backward context codes are essential for economy, even though they destroy the purely word by word character of the coding.

It is much less widely known that one can conceive, for uncorrelated languages, of spellings in which some words are identical to the beginning of others, but any message can still be cut into words by using up to the whole future of the message. In the corresponding coding trees, the vertices representing words are no longer necessarily terminal vertices. The existence of such coding methods was pointed out by Sardinas and Patterson (1953). If the method is such that the whole future is always necessary for the identification of any single word, then the individuality of the words is no more preserved than if the message were frankly coded en-bloc, and such coding methods do not seem worth considering. But if the future necessary to identify a word is finite, with positive probability, one must investigate closer whether the efficiency of coding could be improved by non-recurrence.

Szilard's inequality is no longer true by inspection, but if it is satisfied, the non-recurrent method of coding may be immediately replaced by an equivalent recurrent method, (leading to the same cost of coding). Now, the necessity of Szilard's inequality may be proven by reduction to the absurd. If $\Sigma S(C)M^{-C} = M^{-F}$, with $F < 0$, let us consider the language with word probabilities $p_r = M^{-C_r + F}$. For this language, the cost of coding would be smaller than the information, which contradicts Shannon's theorem on the noiseless channel. (We do not need to try to avoid here the argument based upon the permutations of words.) Therefore, all non-recurrent codes can, in fact, be essentially reduced to recurrent ones, without loss of efficiency.

2. On Limitation of Errors in Transmission in the Presence of Infinitesimal Noise.

The essential conclusion of the preceding discussion is that for every admissible code system for which Szilard's relation is an equality, there exists an uncorrelated language, called reciprocally matched

to, or conjugate to the code system, such that the coded sequence is uncorrelated as well. (It is not, however, an absolutely arbitrary sequence of the letters, because of the condition that any message ends at the end of a word.) These languages are themselves fully characterized by the structure function S(C) to which they are matched, and by the number M.

To compare different possible structure functions, we wish to introduce a new class of problems (apparently not studied in the literature so far, though these, and not the noiseless channel problems, are the realistic simplification of the noisy channel case). Let us study communication in the presence of a very small noise.

The results of noisy channel theory are still true, of course, but also unnecessary, since they are already implied by the condition of small noise. However, there is in practice a great difficulty in the fact that when an error occurs, (and Shannon has shown that conceptually the probability of this happening may be made as small as one wishes, but never zero, for finite messages) the error destroys everything. We wish to add the condition of error limitation, that is, that when an error occurs, it destroys only a limited part of the message. This can be done in two ways only, corresponding to the naively obvious ways of recurrent coding. These coding methods are in a sense the only ones to fully realize the word-by-word character of the coding.

In the first, all word codes have same length. The event: end of a word, is a sure periodic event. (Remark that this method ceases to be error limiting if one admits the possibility that a letter may simply drop out instead of being mistransmitted.) The structure function is zero, except for one value $C=C_o$, for which it is M^{C_o}. The conjugate language has equiprobable words, of probability M^{-C_o}.

In the second method, the M letters divide between M' inside letters, and M-M' "spaces", which may be indifferently considered as initial or terminal letters. If the coding tree is complete, the number of words is infinite, and a fixed percentage of the vertices at each distance from the root are terminal vertices, and represent words. Then $S(C) = (M-M')M'^{C-1}$. If $M' > 1$ the rank function, or number of words of length less than C, is

$$r(C) = (M-M')(M'^{C}-1)/(M'-1)$$

The probability of the word of length C being M^{-C}, it may be written, for large r, as

$$p_r = Pr^{-B} \text{ , where } B = (\log M)/(\log M') > 1.$$

If $M' = 1$, $r(C) = C(M-M')$, $p_r = C^{-B}r$, where $B = \log_e M/(M-M')$, one can combine the two methods above by deciding that if no letter before the L^{th} is a space, the word ends on the L^{th} letter, whichever it is.

In the last two methods, if a letter is mistransmitted, one word

is destroyed; if a space is mistransmitted as letter, two words are destroyed.

3. Noise Limiting Coding Problem in the Absence of Matching Between the Language and the Alphabet.

3.1. The Problem. In the realistic case, the choice of the words of the language is not fully imposed a priori. The language presents itself as a sequence of elements E_r, either uncorrelated but not equiprobable (let their probabilities be p_r), or forming a Markoff chain (let the transition probabilities be p_{ij}). The alphabet can be given as a collection of elements L_g, to each of which is attached a sort of cost C_g; which may be made to depend upon the preceding letter in the coded sequence, in which case, we call them C_{ij}. One wishes to represent the elements of the message, or combinations of these elements decided upon in advance, through letters or combinations of letters, with the least resultant cost of coding per element. The combinations of elements used as words can be chosen with this aim in view. Or else, the alphabet is a collection of elements L_g, which one wishes, for matching, to use with well determined probabilities p_g, or with well-determined transition probabilities p_{ij}. In the case of independence, to give the C_g is equivalent to giving the p_g, since one can choose one, and only one, C, such that $C_g = -C \log_e p_g$, where $\Sigma p_g = 1$. (It is sufficient to solve $\sum_g e^{-C_g/C} = 1$. But this does not work in the case of Markoff dependence, since one should have $\Sigma_j p_{ij} = 1$, and there is no reason, in general, for all C_i, obtained in the above way, to be equal. Therefore, to give the C_{ij} is more general than to give the p_{ij}.

3.2. Shannon's Coding Theory. Shannon considers words each composed of the same large number N of elements of the language. By the law of large numbers, these words divide into a class containing 2^{NH} words of probability close to 2^{-NH} (by definition of $H = -\Sigma p_i p_{ij} \log p_{ij}$, where p_i are the absolute probabilities, deducible from the p_{ij}), and a second class containing words of total probability tending to zero with N. (Precise results are given by Shannon, p. 397.)

Similarly, when one takes strings of a fixed number N' of letters, these strings, except for an arbitrary small percentage, will have costs equal to and independent from the preceding string. This defines the capacity of the alphabet.

To code the arbitrary language by the arbitrary alphabet, it is then sufficient to match the lengths of the strings of elements of the language and of the letters of the alphabet, so that their ratio is equal to the ratio of information to capacity. The words of the low frequency class are never transmitted and the sequences of letters not having the average properties are never used.

The only reason for the success of Shannon's coding method is that, through the law of large numbers, one has "generated" from the initial elements new elements, the structure function of which is sim-

plified and independent from that of the generating elements. This structure function is that of the first case which can be coded by random sequences of letters with error limitation. This procedure of smoothing out structure functions by addition is, of course, identical to that used in statistical mechanics to explain the regularity of observable macroscopic quantities, while making on their components very slight hypotheses only.

3.3. Proposal for an Alternative Coding Theory. We want now to draw attention to the fact that these reasons for the success of Shannon's coding theory fully subsist if the compound words are generated by the second of the error-limiting methods. One chooses one element to play the role of space, and defines words as being all sequences of the initial elements between two space symbols. The theory based upon this generation of the words is conceptually parallel to Shannon's theory, but it is more useful for the description of the most important single class of statistical languages: the natural languages.

The essential result is the following combinatorial theorem, which is proved in the appendix. Consider a discrete finite irreducible Markoff chain. Instead of cutting it into stretches from the outside, let it cut itself, by specializing one of the states to be spaces. The number of letters other than space must be > 1. Define words to be all possible sequences of letters between two successive spaces. Words are uncorrelated. The number of words R is infinite. The structure function S(C) of these words is strongly simplified and independent of that of the generating elements: one shows that it is an exponential $S(C) = M_0{}^C$, like in the equiprobable, independent case. If words are ranked by decreasing probability, their distribution law is therefore again $p_r = P\, r^{-B}$, where B is a fairly complex function (> 1) of the initial Markoff process. This formula is not valid for r small.

The number of generated words being infinite, no concept of information can appear as logarithm of the number of generated words. But the role of H as measure of number of words, of "wealth of vocabulary", is now played by B, or rather by $1/(B-1)$.

Similarly, consider an arbitrary coding alphabet which one wishes to see used in the coded message with the probabilities q_{ij} (the number of letters does not have to be equal to the number of elements of the language), and also in the second error-limiting fashion (all codes for words end by space). But the same theorem as above still applies; our requirement is equivalent to wishing to use the coding sequences of letters with the probabilities $p_r = P\, r^{-B}$.

Let us now suppose that the choices of the language and alphabet spaces have been made so that the two corresponding B's are equal. The coding procedure codes each generated word by the letter sequence of same rank. Clearly, the coded sequence has the desired statistical properties. The only difficulties could come from the most

frequent words and letter sequences. But these difficulties are arbitrarily reduced if the probability of returning from space to space in, say, three steps is supposed small enough, which requires B to be very close to 1.

(If the channel is not given by the desired probabilities of the letters, but by their costs, one must rank all sequences of letters ending by space, by order of increasing cost. This is the same as the problem of ranking of words by decreasing probabilities, except that $-\log p_{ij}$ is replaced by C_{ij}. The result is $C_r = C' + C'' \log r$. The coding method again puts together words and spellings of same rank.)

In Shannon's theory of coding and in ours, the generated words are coded in an error-limiting fashion. There are presumably many other systematic ways of generating words which smooth out the properties of the generating elements, but the coding methods cannot be error-limiting.

By definition, we shall refer to the law $p_r = P\, r^{-B}$ as the canonical distribution law.

3.4. Maximum Information Properties of Generated Languages. Both Shannon's generated languages and ours possess interesting external properties. Let cost of coding be measured in terms of $C_{ij} = -\log p_{ij}$. Then, Shannon's language with equiprobable words, transmits the maximum of information compatible with a given maximum cost of coding per word. The canonical languages transmit the maximum of information compatible with a given average cost of coding per word, together with the requirement that a space be used.

3.5. Generalization of Languages Having the Maximum Information Property. But if, inversely, we take this maximum information property as a criterion, instead of obtaining it as a theorem, we find other laws beside the one above. ($R = \infty$, $B > 1$) The formal appearance is still the same: $p_r = P\, r^{-B}$, but B does not need to be > 1, and R need not be infinite (it must even be finite if $B < 1$). Therefore, the canonical family can be prolonged, to include even the equiprobable distribution (case $B = 0$).

To obtain the maximum information we must first of all attribute the cheapest code to the most frequent word, etc., the cost increasing when the probability decreases. Therefore, the relationship between cost and rank obtained for the purposes of the generation of words is still valid: for $r >> 1$, $C_r = C' + C'' \log r$. We assume $\bar{C} = \Sigma p_r C_r$ to be fixed, and want to maximize $H = -\Sigma p_r \log p_r$. For that purpose, we minimize $C - BH$. We obtain:

$$p_r = P' e^{-B' C_r} = P\, r^{-B}$$

as above. But B is not determined by the alphabet anymore: it is determined by the imposed condition that C take some imposed value, less than $\log R$, which corresponds to $B = 0$. See for details in Mandelbrot (1953)(1954a).

The above criterion is fully identical formally to Boltzmann's criterion for equilibrium of a gas. The only difference is due to the special form of the structure function. We know that it is here an exponential. (This is due to the possibility of discriminating between two words differing only by the order of the letters.) In physics, on the contrary, the structure function never increases faster than a power. The result of this difference is far reaching. In physics, high energy (high C) states are rarely occupied, and the number of states may always be considered as infinite. When $1/B$ tends to 0, the $\bar{C}$ tends to infinity, and whichever the imposed $\bar{C}$, there is a value of B which leads to it. Here however, the high energy states play an excessively important role. If R, the number of states, is finite, B can go down to zero, but $\bar{C}$ stays finite when $1/B$ goes to infinity: therefore, the values of $\bar{C}$ which one may impose on oneself at the outset are limited when one fixes the alphabet and the number of states R. If R is infinite, B must stay > 1, and any value of $\bar{C}$ may be obtained by putting B close enough to 1. Therefore, there is always a restriction, either on R or on B whereas in physics, there is none. (See more details; Mandelbrot, 1954a)

$1/B = 1$ is a sort of critical temperature, dividing two zones of quite different behavior of everything*.

4. Statistical Structure of the Natural Languages.

4.1. Empirical Observation. It has been experimentally found by linguists and psychologists that, if natural languages are considered as sequences of words, a word being anything between two symbols of space (in the natural sense of space), then the absolute probabilities of the natural words follow the law $p_r = P r^{-B}$, which is what we have found above to result from the generation, from arbitrary languages,

* It may be interesting to exhibit a zero redundancy code for the case where $B = 1$; and where R, finite, is such that there exists a k such that $M^{M^k} - 1 = R(M-1)$, where M is the number of letters. The exact form of the language considered (and which for r large, but still necessarily finite, behaves like $P r^{-1}$) is described by the fact that there are M^C words of probability M^{-k-C}, where C goes from 0 to $M^k - 1$. Thus, $-\log_M p = k + C$. This form suggests a coding method in which each spelling contains two parts. The first has a fixed number of letters k, and the number formed by these letters gives the number of letters in the second part of the spelling.

(Such a two-part code could even be used to generate other canonical languages with $B > 1$, if one admits all combinations of the k first letters, but supposes that the following letters are correlated, so that only a number M'^k, instead of M^k, are actually used.)

of languages codable without redundancy in the second noise-limiting fashion. In the great majority of cases, $B > 1$. For data, see Estoup, Zipf, Baker, Josselson (references). Some other authors do not study the distribution of words within one sample, but the variation of the number of different words V as the total length N of the sample in number of words increases. They do find the variation $V = KN^{1/B}$, as follows theoretically from $p_r = Pr^{-B}$ if $B > 1$, and R is infinite; except that for large N, there is some flattening of the curve (V, N), suggesting that R is in fact finite, but does not influence average samples in an appreciable fashion. For data, see Chotlos (1944), Baker (1950).

The number B is the principal "variable of state" of the weighted vocabularies studied in this fashion. 1/B or 1/(B-1) are measures of wealth of vocabulary, more intrinsic than the total available number of words, the influence of which is little felt on average samples.

One unquestionable consequence of this empirical finding is that if an alphabet is arbitrary, except in that it leads to the right value of B, that is, depends upon the persons to whose speech it is matched, then the coding by this alphabet is an uncorrelated sequence of letters. For each speaker, there exists a private unbreakable secret code.*

The matching by equalization of the B's only, is justified when B is > 1 and very close to 1, that is, if the first few words are really very unimportant. If B is not so small, one should be careful to match the first few words also. To avoid going into full details of the alphabet and of the elements of the language, which would destroy the purpose of our smoothing theory, we suggest a way to improve the law $p_r = Pr^{-B}$, by writing it as

$$p_r = P(r + \rho)^{-B} .$$

This formula, exact for some simple kinds of alphabet, seems also to be a good approximation for others. In these new conditions the coding alphabet should be matched to two parameters of the language: B and ρ. (ρ could be replaced by H, which is a function of B and ρ; if one does so, one obtains a description of a language fully parallel to that of a gas by temperature and entropy.)

The values of B found on the samples examined so far range

* Unbreakable within the condition that one does not use the transition probabilities between words. One may however take care of the most conspicuous of these relations without going into a general study. If for example "a" is the space, "aa" is the code for "the ". As "the" never follows itself, "aaa" would never occur, whereas "bbb ","ccc" etc., would occur, and the space could be spotted, the first step in the deciphering. To avoid this, one may never use "bbb ","ccc" within a word, or sometimes repeat "the ", with a small fixed probability. Only the first word's probability being changed, B does not change.

from B = 1.6 for some schizophrenics and some children, to B very close to 1 for very "sophisticated" authors. (If B is close to 1, ρ is usually very small. Besides, B is very difficult to estimate from the data. However, in addition, P is no more an independent state function, but behaves like $1/(B-1)$. One can therefore estimate B by measuring P from the p_r's under the assumption B = 1, then deducing B from this value of P.)

4.2. Discussion. One can wish to draw inductions from the particular statistics of the natural words. Whenever one finds that the normal Laplace-Gauss law is satisfied by some phenomenon, one commonly assumes, at least as a working hypothesis, that this results from the addition of a fixed great number of component influences, each of them comparatively unimportant. We conjecture as well that the reason for the statistics of natural words being as it is, is that the words are "composed" of many elements; kind of letters, of which each carries comparatively little information. (Similarly, whenever a non-purely linguistic system of signs is found to follow the canonical law, one may try the conjecture that it is composed of more elementary elements, forming a chain with a Markoff property, or with some weaker property, which still leads to the same law.) One could think of trying to check whether the natural letters or phonemes are these ideal generating elements of words. But this would be improbable, since the natural letters and phonemes give the same code for each word, not only as it occurs repeatedly in any person's speech - which is necessary for recurrence - but as it occurs in different persons' speech. Therefore, the optimum coding property which the ideal elements must possess could hardly be satisfied by natural letters of phonemes. One must therefore make the weaker assumption that the structure of speech as a sequence of words is influenced by some other coding, higher up in the receiving brain, considered as an optimal terminal information processing machine. These elements could not be reached by direct experience, but one may try to identify their cost, which should vary like $\log r$ and $\log p_r$, to the time it takes to read any given word. The experiments of Howes and Solomon on reading time seem to fully confirm this hypothesis.

To sum up, the "explanation" of the statistics of words by their generation from "letters" and by a maximum information property, does not imply that any permutation of words be an admissable sentence. But it implies three things: first, the correspondence between words and ideas, which each person builds for himself, is to a great extent arbitrary, and there are always alternative ways of expressing ideas; second, that the choice of the words actually chosen is not only governed by the need for this particular word, but also by an unconscious matching of the frequencies of words to their codes somewhere in the circuit of communication; third, that these last codes form a

recurrent optimal error-limiting system of the second kind.

Let us finally remark that the crucial role which appears to be played by the symbol space, and therefore by protection against error, may be considered as completing the role which protection against noise plays in restricting languages to be digital, discrete. This discretion could not, (no more than the statistics of words) be considered as being influenced by some structure of the universe of ideas to be represented by the language. The only "explanation" for it seems to be that languages need to be relayed a very great number of times, and that in non-discrete languages the accumulation of noise at the relays could not be resisted, and that these languages could not fulfill any aim.

A more detailed discussion of these matters is given in Mandelbrot (1954b)(1954c).

Appendix.
Combinatorial Derivation of the Law $p_r = P r^{-B}$.

We want to show that this result, already proved in the text for the case of equicostly and equiprobable uncorrelated letters, still holds for non-equiprobable and correlated letters. The generalization is the same as the one that goes from the central limit theorem for Bernouilli trials to the central limit theorem on Markoff processes. We shall not consider the mathematical refinements necessary in a few places; these will be presented elsewhere.

Assume a finite irreducible Markoff process, of states L_g where $(0 \leqslant g \leqslant G)$, and where L_0 is the symbol space. Let the transition probabilities be p_{ij}. Let us choose M close enough to 1 so that the numbers $C_{ij} = -\log p_{ij}$ can all be considered as integers. Let a word be any sequence of states between two successive spaces. We want to find the relation between the probability of a word and its rank, that is, the number of words of higher probability. We start by the relation between probability and the number of words of this same probability. For this, take $C = -\log p$ as variable, and count all sequences of letters starting and ending by L_0, and containing no L_0 in the middle, such that the C_{ij} add to C. (In the sequence of the C_{ij}, the second index of each letter is the same as the first one of the next, and the first index of the first and the last index of last are 0.) Let this number of sequences be $S_0(C)$. Let also the auxiliary numbers $S_k(C)$ be the numbers of sequences starting with L_0, ending with L_k, and containing no L in their middle.

One sees by exhaustion that between these functions $S_k(C)$, one has the recurrence relations:

$$S_k(C) = \sum_{h \neq 0} S_h (C - C_{hk}) + \delta (C - C_{0k})$$

where δ is Kronecker's function, equal to 1 or 0, depending upon whether its argument is 0 or $\neq 0$.

Define further the generating functions of the structure functions $S_k(C)$

$$G_k(z) = \sum_C z^C S_k(C)$$

They will be given by the system of linear equations:

$$\sum_{h \neq o} z^{C_{hk}} G_h(z) - G_k(z) = -z^{C_{ok}}$$

which is solved by quotients of determinants. The function $G_o(z)$ is the only one that really interests us. Write it as:

$$G_o(z) = \frac{N_o(z)}{D(z)}$$

The denominator has no index as it is the same for all $G_k(z)$. Let us write it fully:

$$D(z) = \begin{vmatrix} z^{C_{11}-1} & z^{C_{12}} & z^{C_{13}} & . & z^{C_{1G}} \\ z^{C_{21}} & z^{C_{22}-1} & z^{C_{23}} & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ z^{C_{G1}} & . & . & . & z^{C_{GG}-1} \end{vmatrix}$$

write $-\log_m z = \theta$. The above equation can be rewritten as:

$$D(\theta) = \begin{vmatrix} p_{11}^{\theta}-1 & p_{12}^{\theta} & p_{13}^{\theta} & . & p_{1G}^{\theta} \\ p_{21}^{\theta} & p_{22}^{\theta}-1 & p_{23}^{\theta} & . & p_{2G}^{\theta} \\ . & . & . & . & . \\ . & . & . & . & . \\ p_{G1}^{\theta} & . & . & . & p_{GG}^{\theta}-1 \end{vmatrix}$$

where the unwritten terms are filled in easily. We shall call $D(\theta) = 0$ the "eigentemperature" equation of the Markoff process, relative to L. It differs from the characteristic equation of the square array p_{ij} $(1 \leqslant i, j \leqslant G)$, by the fact that θ is here an exponent instead of multiplying the p_{ij}. (This equation is besides not invariant by any kind of usual transformations of the p_{ij}, but this is quite all right, because the state L_o does play a quite privileged role.)

The real parts of the roots θ are all less than 1, and the roots z have moduli greater than 1/M. Proof: If θ is a root, there exists a set of x_i, not all 0, such that $x_i = \sum_j x_j p_{ji}^{\theta}$. Take absolute values:

$$|x_i| \leqslant \sum_j |x_j| \, p_{ji}^{Re\theta}$$

Sum with respect to i:

$$\Sigma |x_i| \leqslant \Sigma |x_i| \sum_j p_{ij}^{Re\theta} .$$

This requires that for at least one i, $\Sigma p_{ij}^{Re\theta} \geqslant 1$. But $\sum_j p_{ij} \leqslant 1$. Therefore $Re\theta \leqslant 1$. Besides it cannot be $=1$, since $\sum_j p_{ij} \leqslant 1$, and for at least one i, $\Sigma p_{ij} < 1$, and therefore

$$\Sigma |x_i| > \Sigma |x_i| \, \Sigma p_{ij} .$$

Therefore $Re\theta < 1$, and it follows that $z > 1/M$.

Further, N_o/D represents a Taylor series with non-negative coefficients, and, therefore, at least one of its poles (roots of D) which are closest to the origin is real. Call it $1/M_o$. Let us now limit ourselves to the case where there is no other pole of same modulus, and call this case regular.

Let us now write the rational function $G_o(z)$ as a sum of partial fractions $\Sigma A_s/(1-M_s z)$. Then $S_o(C)$ will be automatically obtained as a sum of terms $\sum_s A_s M_s^C$.

The largest in modulus of the inverses of roots z is M_o. Therefore, for large C, the structure function behaves like M_o^C. It follows that the rank function, the number of words of $-\log_M p$ smaller than C, is asymptotically for $C >> 1$, proportional to

$$M_o^C = M^{C/B} = (M^{-C})^{-1/B} = p^{-1/B}$$

$$B = 1/\theta = (\log M)/(\log M_o) > 1$$

Inverting, we find that asymptotically for large ranks by order of decreasing frequency, the law giving the probability of a word as a function of the rank is:

$$p_r = P r^{-B}$$

Except for B, this is independent from the initial Markoff process, supposed regular.

Because of the fairly complicated form of the eigentemperature equation, we shall give the much simpler form it takes for uncorrelated letters, when p_{ij} depends only upon j. One easily finds that:

$$D(z) = (-1)^G (1 - \Sigma^G 1 \, z^{C_j})$$

$$D(\theta) = (-1)^G (1 - \Sigma^G 1 \, p_j^{\theta})$$

In the case of independence, one can generalize the problem by

assuming that there are several symbols "space". (In the case of dependence, this generalization would have made the words non-independent, but would not have changed the calculations leading to their absolute probabilities.)

Particular Case when G=1. Then the eigentemperature equation has the root $M_0 = 1$, $\theta = 0$, and B would be infinite. The above study breaks down. One sees, however, by a direct study that the rank of a word, by decreasing frequency, is now simply equal to the number of letters it contains minus 1. The most frequent word has probability p_{00}, the next $p_{01}p_{10}$, the word $n^{o}r$,

$$p_r = p_{01}(p_{11})^{r-2}p_{10} = p_{01}p_{11}p_{10}(p_{11})^{r} = P\,e^{-Br}$$

which is quite different from the usual canonical law.

REFERENCES

S. J. Baker 1950, J. Gen. Psycho., 42, p. 25; 1951, ibid, 44, p. 235.

J. Chotlos 1944, Psycho. Monographs, 56, p. 75.

R. M. Fano 1949, MIT Research Lab. Electronics T. R. No. 65.

W. Feller, 1950, Introduction to Probability Theory, J. Wiley, N. Y.

D. A. Huffman, 1952, Proceedings of the IRE, 40 (1952) p. 1098, or 1953, Communication Theory, Butterworths London, p. 102.

H. Josselson, 1953, Russian Word Count, Wayne U. Press, Detroit.

A. A. Khinchin, 1949, Mathematical Foundations of Statistical Mechanics. Dover, N. Y.

B. Mandelbrot, 1953, Communication Theory, London Butterworth, p. 486.

1954a, Transactions of the IRE, Information Theory, No. 3, p. 124.

1954b, Word, 10, p. 1-27.

1954c, Scientific Monthly, to appear in September.

A. Sardinas, Patterson, 1953, Convention Record of the IRE, Part 8: Information theory.

C. E. Shannon, 1948, Bell S. T. J., 27, 379 and 623.

L. Szilard, 1929, Zeitschrift fur Physik, 53, p. 840.

G. K. Zipf, 1949, Human Behaviour and the Principle of Least Effort, Addidon Wesley Press, Cambridge.