36

AN INFORMATIONAL THEORY OF THE STATISTICAL STRUCTURE OF LANGUAGE

Benoît Mandelbrot
Laboratoires d'Electronique et de Physique Appliquées, Paris

SUMMARY

The problem of the statistical structure of a language best adapted to a word coding is investigated. It is shown that this problem is an inverse of the now classical problem of C. E. Shannon, that is to say, of the 'direct' problem of constructing the least costly coding for a given message.

It is first shown that the physical properties of a transmission channel relevant to the coding can be described by a set of 'macroscopic variables of state', which also characterize the best word-by-word coding possible on this channel.

To express adaptation of the statistical structure of the language to this coding, several criteria of optimality are written down. It turns out that they all lead to a single law, and the corresponding 'canonical' statistics are also described by macroscopic 'variables of state'.

These statistics give a very accurate description of the experimental data on words in actual written languages. In other terms, a quite general statistical structure, entirely independent of meaning, appears, underlying meaningful written languages. This fact is to be considered as a very strong argument in favour of a thesis that language is a message intentionally if not consciously produced in order to be decoded word-by-word in the easiest possible fashion.

SCOPE OF THE THEORY

The aim of the theory given below is to provide an information theoretical model, through which the actual structure of language can be linked to presumed, entirely plausible, properties of the brain. It is hoped that this work will provide the beginning of a mathematical approach in linguistics (as opposed in particular to the purely arithmetical approach of the statisticians) and thereby give a quantitative justification to the assumption that the Theory of Communication, and in particular its economical principles, apply also to the higher function of the living being.

Our theory is based upon the consideration of a single speaker and a single receiver, who is best visualized as the same person in a different capacity, as one may presume that one encodes in the way in which one would wish to receive.

Three elements are to be considered:

(1) The structure of language, or shortly, message.

(2) The way in which information is coded by the brain.
(3) The economical 'criterion of matching' which links 1 and 2.

On the basis of the theory of communication, any one of the above three elements could be deduced from the remaining two, if these were known exactly:

(*a*) If language were the unknown, one would look for the best structure to be given to an artificially constructed language, if it is to satisfy the economical criterion with the codings which have been assumed.

(*b*) If the criterion were the unknown, one would code the actual languages in the fashion corresponding to the actual coding principle, supposed known, and then look for the properties common to all the sequences of signs thus obtained.

(*c*) If coding were the unknown, to be chosen from the limited number of possible codings, one would check all these codings exhaustively.

Unfortunately, none of the elements is given in a form directly suitable for use in a mathematical theory, but must first be extracted by qualitative discussion and analysis from all the obtainable data, historical and structural, on languages, and eventually in other fields. Such a qualitative discussion cannot give us more than a reasonable degree of confidence in the results to which it leads. Therefore the aim of a mathematical theory cannot be, strictly speaking, to deduce one of the elements from the remaining two, but only to add (very considerably) to the confidence we put in each of these elements individually, by showing that they are fully compatible.

However, from a technical (as opposed to conceptual) point of view, the proof of the compatibility cannot avoid the use of one of the above-mentioned deductive methods.

It will turn out later that the easiest is the first method, *i.e.* language as the unknown, (which incidentally goes from some kind of overall psychophysiology to linguistics, that is, goes up on the classification of the sciences of Auguste Comte*). This method is in particular the only one which can be carried through mathematically, without requiring a preliminary numerical representation of the structure of language, which would in particular contain the extremely difficult problem of the division into elements of a continuous string of sounds. (By 'divide' we mean both divide in principle and design the apparatus to do the job.)

This difficulty of representation arises in all the cases where one has to describe something without a comparison set. Then the easiest method is not the 'direct' one, but usually the 'inverse method': one builds up a privileged set of labelled messages, and tries to identify the given message with one of the set; the division into elements is obtained *a posteriori*.

This method also gives an opportunity to try successively several variants of the economical criterion of optimality, justified by a principle of 'matching' through evolutionary optimization. Finally, the details of the coding will themselves turn out to be irrelevant, as its specification will be replaced by a set of 'state variables'.

* See Appendix p. 499.

PLAN OF THE PAPER

The first part of the paper is devoted to a sketch of the theory of language on which our theory will rely, and which is due to the great Swiss linguist FERDINAND DE SAUSSURE (1857–1913)[1]. It is a striking and remarkable fact that his theory can be exposed in the context of modern Communication Theory with very slight alterations only. None of the modern qualitative concepts would have been unfamiliar to de Saussure, who states explicitly that language is just another means of transmitting information, and as such should be studied as one aspect of 'semiology', or science of signs. de Saussure could be considered as the founder of the abstract and general Theory of Communication, and the words 'Semiology', or 'Semiometry' might still be good alternative choices for 'Theory of Communication'.

de Saussure stressed the peculiar features of language among other signs, features that make its study the prototype of all other semiological studies. Among these is the arbitrariness, or rather unmotivated character of the linguistic sign. However, he went too far in this direction, and considered language as the only means of transmitting information in which there is no necessity for adaptation of the message and its support. In this respect only we wholly disagree with de Saussure; our theory is based on the exactly opposite assumption that language requires such a matching, and even is bound to show its effects in the most apparent and quantitative fashion. This only discrepancy is essential, as it is what makes possible a continuation of de Saussure's theory in the second part of the paper.

This part will be devoted to purely information-theoretical quantitative considerations. These will show the power both of the conceptual framework of de Saussure and of the technical tools of communication theory. A single 'canonical' family of statistical distribution laws will be obtained, which has some very important features, and accounts for all the data on words collected by ZIPF[2]. This will confirm quantitatively de Saussure's assumption that language can be considered as built out of a sequence of words.

LINGUISTIC: MESSAGE, CODE AND CRITERION WITHIN THE FRAMEWORK OF DE SAUSSURE'S STRUCTURAL AND FUNCTIONAL THEORY OF LANGUAGE

Concrete entities or units of language

In the actually observed structure of the language, there are several levels of 'natural' signs—letters or phonemes—then groups of these and in particular words, then groups of words. Our present purpose is to discuss the relations of these various levels with the presumed properties of the brain.

Whatever the detailed structure of the brain, it recodes information many times. The public representation through phonemes is immediately replaced by a private one through a string of nerve impulses. (These recodings may become quite conscious, at least differentially, as one switches from 'right' to 'write'.)

This recoded message presumably uses fewer signs than the incoming one; therefore, when a given message reaches a higher level, it will have been

reduced to a choice between a few possibilities only, without the extreme redundancy of the sounds. The last stages are 'idea' stages, where not only the public representation has been lost, but also the public elements of information.

The message will be said to be matched to a decoding process, D, at a certain level if, in some meaning to be made more precise later, the 'work of decoding' by this D of the receiver is the smallest possible. Matching will relate a definition of 'sign' and a statistical structure to the corresponding coding, (see p. 492 onwards).

The most obvious and elementary structure is that of phonemes, letters or sounds. These are the only visible coding stages, and, from the point of view of the economy of the whole, the least critical ones, as they are conditioned by muscle and not nerve. Moreover, they are greatly influenced by historical accidents and the rigidity of 'vested interests' (like the English monetary and measuring system). Therefore, there was *a priori* little hope of finding that the matching exceeds the familiar approximate relation between length and rarity. This is incidentally very fortunate in a way, because matching achieved higher up is bound to give information on more interesting aspects of the living being.

For the best results, matching should prevail at all levels, but most necessarily at the highest ones. One would therefore strive to reach the highest of the 'idea' levels. But such a matching would be heavily dependent upon the person considered, and can be assumed only in the case of a single reader or listener (except for the reciprocal matching of the 'semantical' level to frequently received ideas, upon which are based all commercial and political propaganda). Even in the best case, this level is difficult to estimate from the outside and to match. Matching could be only approximate and is difficult to study in a rational physical fashion.

Moreover, the idea and the letter levels would not be part of linguistics as understood by de Saussure, which is our present business. The object of that science is in fact the study of one particular level, which is what remains as consequence of a 'grinding' process in which individual 'acts of speech' get averaged into all acts of speech of a single speaker, then into all acts of all speakers*. This grinding and averaging process which leads to the formation of a common language as a tool is at the same time the one which should be followed in excluding from the study of language as an object of science all the 'irrelevant and extraneous elements of the complex whole, which is given to our common experience'. In particular, this grinding process should give the elements out of which language is built. These, and their statistical structure, should be the same for the listener, whoever the speaker, and conversely.

In short, pushing de Saussure's theory to its logical limit, our theory will consider language as a random sequence of concrete entities, and will be carried to an estimation of the probabilities of concrete entities. This is a

* DE SAUSSURE notes himself that the job of the linguist is similar to that of the man trying to find out about the essentials of the game of chess by considering first a single game, then a set of games by one man, and finally all games. This assimilation of language to a game is very deep, as it will turn out that the tool which will make possible a mathematical study of the decoding process will be the modern theory of games of strategy as applied to one of the aspects of de Saussure's analogy.

tremendous simplification of language, just as extreme, but exactly opposite to, the one which leads to the Semantics of the Vienna School (SCHLICK, REICHENBACH, CARNAP).

Digital character of the coding

The two main *modes* of representation, already recognized by de Saussure, are respectively imitative (analog) and symbolic (digital). There are two good reasons to believe that the essential codings are digital.

Imitative elements are concrete and therefore susceptible of continuous variations, and each idea requires a different sign. As these must be few in order to be distinguished, a purely imitative language could not serve as a civilized language. These were made possible only by the invention of the possibility of the synthesis of complex ideas by combination of simpler ones. But such a combination amplifies the consequences of small local errors, and therefore its development had to be accompanied by a progressive standardization of the elementary ideas themselves, or in other terms by the invention of some kind of concrete entity. To express ideas by combinations of standard words is a typically digital method of coding. (It would be extremely interesting to follow, in the spirit of the theory of coding, the progressive breaking up of phrases into words by young children.)

If further recodings of a digital coding were not themselves digital, there would be no sense in the first coding being digital, at the expense of so much trouble. But in fact, this is obviously the case for spelling, the only representation which has been consciously evolved for this purpose, the history of which is a long effort from imitative to symbolic coding. On the other side, the unconscious coding by 'sounds' does not seem to be really digital. But it has later been recognized that sounds are not perceived as such, but as 'phonemes', which are 'classes of equivalence' of sounds. The phonematic system of a language is based on a certain number of 'oppositions', which are simply coding dichotomies; therefore, despite the appearances, the coding system of the spoken language is also typically digital.

The same result would be obtained from the neurophysiological evidence that the functioning of the neurones is digital. But one would have to admit that the coding is based on neurones at the critical stage, which is far from being evident.

Least effort character of the matching criterion

This is based mainly on evidence diffuse in a large number of other fields, since, as Professor A. S. C. Ross pointed out in the discussion, one had better avoid any conclusion from a hypothetical history of language.

Macroscopic description of a text

The results of the preceding discussions are independent of the language considered. Therefore the quantitative results on the statistical structure of the concrete entities, which are to be used as a criterion for their identification among the various units of information, must also be the same for all languages.

In fact, the mathematical details will show that all the variants of the criterion of least effort lead to the same 'canonical' family of laws for concrete entities. Let us rank them in the order of decreasing frequency. Then, the frequency p_n of the nth entity in this classification should be given by

$$p_n = P(n + m)^{-B}$$

where P, m, B are certain positive constants, to be derived below.

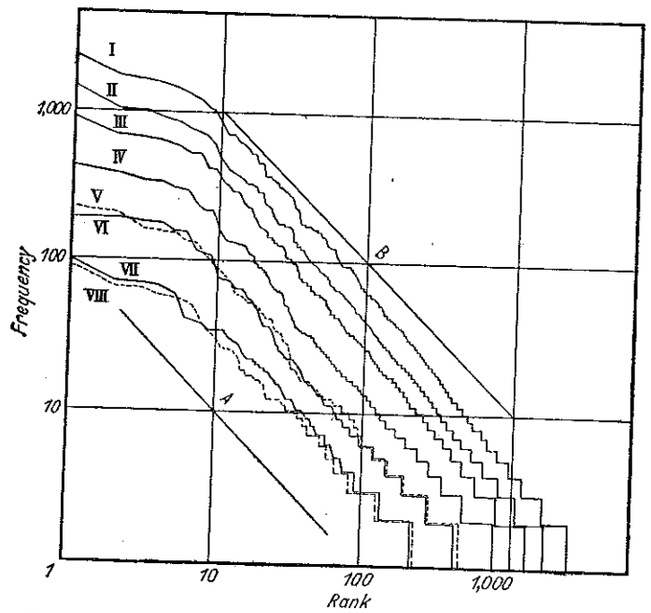

Figure 1. Rank-frequency distributions of Helen B. (with paranoid schizophrenia) in samples of (I) 50,000 words; (II) 30,000 words; (III) 20,000 words; (IV) 10,000 words; (V and VI) 5,000 words, and (VII and VIII) 2,000 words. (From *Arch. Neurol. Psychiat.* 49 (1943) 831)

It turns out that this formula, shown in *Figure 2*, represents accurately all the experimental data on the statistics of words, if these are taken in fully inflected form, and not as lexical units of the dictionary. Therefore words can be considered as being concrete entities, though they may not be the only ones. The best check is obtained where semantic constraints are the least, that is in schizophrenics (*Figure 1*).

In the cases where 'word' is not unambiguously defined, our theory cannot be considered as a proof of compatibility of three known elements, as was said on p. 487. But it may be used as a discriminating rule to choose the better of several possible divisions of the string of phonemes into words.

The definition of word as a unit of the dictionary was used chiefly by UDNY YULE. It leads to different results, depending not only on language used, but also on author, and is therefore appropriate for the search of the paternity of literary texts (which was the purpose of YULE's study).

Our own definition was used by ZIPF. In fact, the above formula is very similar to one introduced empirically by J. B. ESTOUP and studied by Zipf, and which is the particular case of ours for $B = 1$, $m = 0$, namely

$$p_n = Pn^{-1}$$

Our research has originated from this formula. But Zipf's exposition was not based in the least on any serious linguistic theory, nor on communication

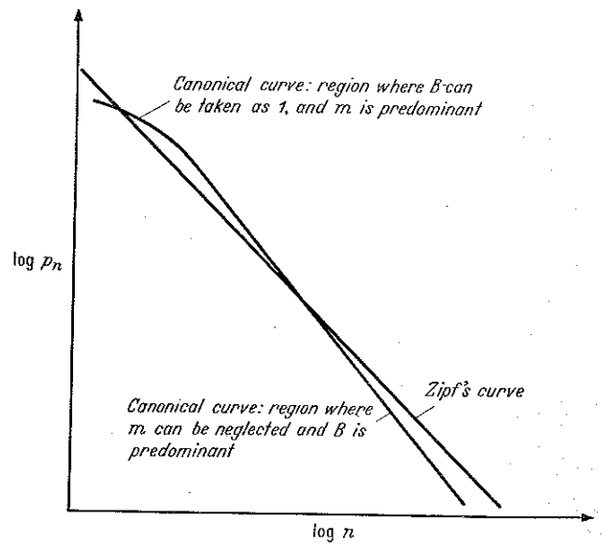

Figure 2. Canonical Rank-frequency distribution $p_n = P(n + m)^{-B}$ as compared with Zipf's "law" $pn = Pn^{-1}$

theory. Moreover, the new formula adds two parameters, m and B, which constitute necessary improvements, as they account for quite marked discrepancies from the general trend, represented by Zipf's law: the first improvement is more important for n small, and the second for n large.

Their striking character is due to the fact that they are of mathematical, and not of empirical, origin. Their importance shows that this theory exhibits a fascinating feature of thermodynamics; namely 'every parameter of mathematical origin receives almost inevitably a fundamental physical meaning' (SCHRÖDINGER). We shall take up this point in more detail later.

MATHEMATICAL: MICROSCOPIC STUDY OF THE STATISTICAL MATCHING BETWEEN CODING AND MESSAGE

Two reciprocal forms of matching

This is a study of a formal relation which can be satisfied by any digital message and coding, irrespective of the actual meaning of the term 'word'. It was found very helpful to consider this relation in two independent and opposite fashions, depending upon which term is considered as fixed. (Great

care should be taken not to confuse this distinction with that between encoding and decoding.)

The direct problem of statistical matching is the now classical problem of SHANNON, namely to find the least costly method of coding for a given message. The solution of this problem is a many-to-one mapping from message space to coding space.

The inverse problem of statistical matching is (the coding having been given) to distribute information among the available 'words', in such a fashion that the capabilities of the coding are utilized fully, following a certain criterion, which in most cases will be an economy criterion. The solution of this problem is given by a many-to-one or many-to-many mapping of coding space to message space. This problem is technically far easier than the first, even if one takes account of the simple direct problem which must be solved to construct the reference coding process. We shall show that it is physically very realistic, since it leads to the actual distribution of information among words of the language. The use of this method introduces a methodological difference between this and other papers of the symposium.

A simple direct problem

Our coding will preserve the individuality of the concrete entities, as this makes the problem simpler.

Let the available elementary symbols S_g, $(1 \leqslant g \leqslant q)$ have costs C_g*. This concept of 'cost' includes everything and anything which enters into the expense of sending S_g, properly weighted. It does not require the concept of 'digit'; on the contrary it leads to it. There is no danger, however, in visualizing C_g as a cost in digits.

In the solution of the direct problem with the restriction of word-by-word coding, the nth word by order of decreasing frequency must be represented by the nth sequence of signs by order of increasing cost. Thus the solution of our direct problem depends only on this ranking, that is on the relative values of these elementary costs, or in other terms on the physical properties of the transmission channel. It can be shown that the cost of the nth sequence can be written in the form

$$C_n = [\log_M (n + m) + j_0] \qquad \ldots . (1)$$

where a number in brackets means the next higher integer to it, and j_0, M and m are a set of constants which are functions of the C_g's and which may be called the macroscopic variables of state of the coding or of the channel, they replace its full specification, and different codings with the same variables of state are to be considered as fully equivalent for what follows.

The above formula is only approximate for more complex forms of C_g, but strictly true for simpler ones. For example, if for $1 \leqslant g \leqslant q$, all $C_g = 1$, the number of sequences of length equal to C is q^C, and the number of sequence of length $\leqslant C$ is $q(q-1)^{-1}(q^C - 1)$. Therefore

$$C_n = \left[-1 + \frac{\log (q-1)}{\log q} + \log_q \left(n + \frac{q}{q-1} \right) \right] \qquad \ldots . (2)$$

* C_g, and later p_g, will be used for the cost and probability of elementary symbols; C_n, and later p_n, for the cost and probability of words.

i.e. $M = q;\ j_0 = -1 + \log(q-1)/\log q;\ m = q\,.\,(q-1)^{-1}$*

It will turn out that the square brackets vanish, and j_0 is not an important constant. But m is an essential one; it is unfortunately often forgotten, though it reveals an extremely characteristic feature of digital coding.

However, word-by-word coding is not the most efficient of the representations permitting the reconstitution of the initial message. SHANNON has shown that one should actually code by large blocks of words. At the limit of representation of the whole message as one single unit, the average cost of the word would be reduced to $H = -\Sigma p_n \log p_n$. This expression gives in a sense a 'value' of the message, and is called its selective information.

It gives also a measure of the surprise value of a word. There are also other kinds of information, semantic or otherwise, and all of them together may still fail to give a complete picture of the totality of language. But, instead of discussing their merits *a priori*, we shall investigate which properties language should have, if its function were only to carry this selective information.

These preliminaries would not be necessary if one wanted only to know which values should take the probabilities p_g of the symbols S_g of cost C_g if the information per average cost is to be the largest possible. To solve this, one minimizes

$$\frac{C}{H} = \Sigma p_g C_g [-\Sigma p_g \log p_g]^{-1}$$

which gives

$$p_g = A e^{-BC_g} \qquad \ldots (3)$$

B must be equal to H/C, which leads to the characteristic equation

$$A^{-1} = \sum_1^q e^{-BC_g} = 1 \quad \text{or} \quad \sum_1^q M^{-C_g} = 1, \text{ with } M = e^{-B}$$

Then the average cost per unit of information becomes

$$C/H = (\log M)^{-1}$$

For this to be a minimum, M must be the largest root of the characteristic equation, which is the same number as in the beginning of this section.

The inverse problem

The above 'best' probabilities for C_g would not be attained with the above 'best' coding word-by-word, except in special cases. This optimum requires the coding with infinite delay, which is never possible.

The main problem in practice is to see how it is possible to compromise between the advantage of 'word-by-word' coding from the point of view of delay, and of 'large block' coding from the point of view of economy.

This does not involve Shannon's problem, but an inverse problem, the distribution of the information among the R different words M_n available. R will be called the potential number of different words. Their full utilization may be understood in several fashions.

* If $C_g = g$ and q is large, $m = 1$. This may give a model explaining HICK's formula for the delay T in the choice between n alternatives, $T = \log(n+1)$, better than Hick's conjecture that there is an additional alternative 'no signal' equiprobable with the n others.

Criteria and corresponding solutions—(*a*) One wants to minimize the excess cost between the coding word per word (where one kind of symbol, S_0, must be reserved for tops) and the Shannon type of coding where all kinds of symbols can be used for information and the coding can be done in blocks of any size.

Let C_0 be the cost of S_0. Equation 3 must then be replaced by

$$p'_g = M'^{-C_g}$$

where g varies from 0 to q instead of from 1 to q, and M' is the largest root of

$$\sum_0^q M'^{-C_g} = 1 \text{ (whereas } \sum_1^q M^{-C_g} = 1)$$

The average cost per unit of information would be only $C'/H = (\log M')^{-1}$ Therefore, the excess cost is

$$(\log M)^{-1} \Sigma p_n C_n + (\log M')^{-1} \Sigma p_n \log'_M p_n$$

This is minimum when p_n assumes the form

$$p_n = P M^{-BC_n} \qquad \ldots (4)$$

where

$$B = \frac{\log M'}{\log M} \geqslant 1$$

B is determined by the transmission channel and R is a free variable, *i.e.* it can be chosen so as to give any contemplated H with the imposed value of B (*Figure 3*), as long as H is less than $H(B, \infty)$.

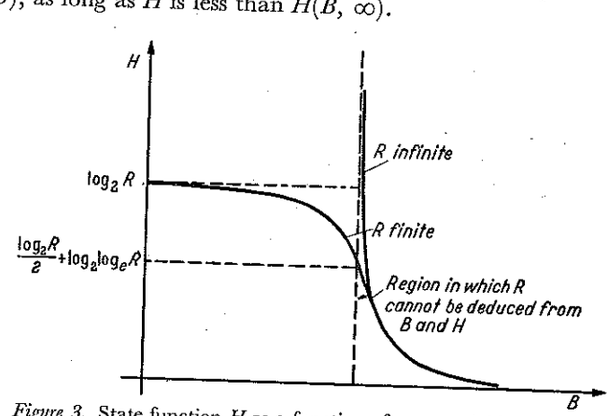

Figure 3. State function H as a function of state variables B and R

(*b*) Our coding represents the words (which are elements of a stochastic process) by groups of more elementary signs. In general, these signs do not remain random. Let us see which values of p_n would keep them so, the tops being included among the signs.

If the word number n requires f symbols, its probability must be

$$p_n = (1-p'_0)^{f-1} p'_0 \Pi \frac{p'_g}{1-p'_0} = \frac{p'_0}{1-p'_0} \Pi p'_g = \frac{p'_0}{1-p'_0} M'^{-C_n}$$

Therefore, p_n assumes the same form as in (*a*) above

$$p_n = \frac{p'_0}{1 - p'_0} M^{-BC_n} \text{ where } B = \frac{\log M'}{\log M} \text{ is also the same number.}$$

But here, the potential number of different words, R, is not free, but determined by the coding and must be infinite. Therefore H is also imposed by the channel and the chosen criterion.

(The same result can also be found, independently of the values of p'_g obtained above, by requiring that p_n depend only upon C_n, which is the only quantity to have an intrinsic physical meaning. Then, p_g being multiplicative, and C_g additive, p'_g must be an exponential of C_g, with no multiplicative constant.)

(*c*) For the criterion of complete prevision one leaves the information undetermined, and wants to minimize the average cost (top included) per unit of information.

$$(C + C_0)H^{-1}$$

is minimum when

$$(\Sigma \Delta p_n C_n)H - (\Sigma \Delta p_n \log p_n)\,(C + C_0) - K^2 H^2 \Sigma \Delta p_n = 0$$

which gives the same p_n, with R again free, and B determined by both the chosen value of R and the coding.

(*d*) In considering the criterion of prevision limited to one word, one has decided upon the quantity of information H to be transmitted and wants to choose the probabilities so as to minimize the average cost C. The method of Lagrangian multipliers leads to the minimization of $C - B^{-1}H$; therefore, one gets again the previous p_n. Now B is free and can be determined from H (*Figure 3*), as long as H is less than $\log R$.

Resulting canonical law

Let us put the value of C_n given by the direct problem into the fundamental law of equation 4 to which the above various criteria have led. We obtain

$$p_n = P[n + m]_M^{-B} \qquad \ldots\ldots(5)$$

where $[x]_M$ is the number such that

$$\log_M [x]_M = [\log_M x + j_0]$$

In a series of actual events, the law of Poisson introduces a dispersion of the frequencies on both sides of the probability, and the steps of the above curve smooth out, to give the 'canonical law'

$$p_n = P(n + m)^{-B} \text{ where } P^{-1} = \sum_1^R (n + m)^{-B} \qquad \ldots\ldots(6)$$

If this law is satisfied, and depending upon the values of R and B, at least one of the criteria in the 'Inverse Problem' is satisfied.

Remark on best value of M

The number M has, as we have seen, little influence upon the actual curve of frequencies. But let it be chosen at random, and the 'letters' be equiprobable; then in general the frequency of the symbol 'top', inverse of the average

length $C(M)$, is different from that of the letters since it depends also upon the information. Therefore the top can be recognized as such, unless $(M + 1) = C(M)$. As $C(M) \log_2 M = C(2)$ (approximately), M must satisfy the equation $(M + 1) \log_2 M = C(2)$. For English, this equation gives $M = 4{\cdot}5$, which is also the value of the actual average length. The replacement of ideal letters by the actual spelling is therefore an expansion in the number of different signs but not in the total number of signs. The redundancy is $1 - (\log 4{\cdot}5/\log 26) = 54$ per cent which checks excellently with Shannon's experimental results.

Functional and thermodynamic interpretation of the canonical law

One may consider the decoder as a physical piece of apparatus, and want to minimize the irreversibility corresponding necessarily to the process of decoding. This point of view does not bring in any new result, but provides the possibility of interpreting the preceding results, and makes clearer the role of selective information.

Let us visualize the process of transmission as a sequence of processes (D) and (D^{-1}).

(D) cuts a continuous incoming string of signs into groups, and recodes each group separately. We shall call it a differentiator, but it is also an identifier and a multiplier, and it requires a memory. It is the counterpart of VON NEUMANN'S process 1 of Quantum Mechanics[3] and of the process of separation of the elements of a mixture[4]. It cuts the incoming string of signs in groups, identifies each group without destroying it and represents it by a new code. By this process (D) the total physical entropy must increase by an amount, which we shall call E_n, when the word M_n comes out. Let us write it under the form

$$E_n = -k \log_e q_n + E_0 \text{ where } \Sigma q_n = 1$$

and therefore

$$E_0 = -\log_e \Sigma \exp(-E_n/k)$$

q_n will be the pseudo-probability of (M_n).

(D^{-1}) is the process inverse to (D). We shall call it integrator. It reconstitutes a new message out of the individual words. This represents a decrease of entropy by an amount

$$k\Sigma p_n \log p_n$$

known as the entropy of mixing[5].

Therefore, the sequence of (D) and (D^{-1}) provides an average increase of entropy by the amount

$$k\Sigma p_n \log\left(\frac{q_n}{p_n}\right) + E_0$$

This is a minimum and equal to E_0 when $p_n = q_n$, which is an ergodic relation, as it expresses the equality of an average upon time to an average upon a piece of machinery. As this minimum must still be positive, we must have Szilard's inequality[4]

$$\Sigma \exp(-E_n/k) \leqslant 1$$

(This inequality could of course be written $E_n \geqslant -k \log q_n$ which seems sharper, because it is valid for every part-process, but actually q_n requires all the other processes for its definition.)

Thus a lower limit to E_n is imposed by thermodynamics. To design a (D) with given q_n is a problem of research of a statistical decision function, (D) being a strategy in the sense of BOREL[6], VON NEUMANN[7], and WALD[8]. The direct problem has usually no solution for arbitrary q_n, but this is not important, since the matching of (D) and messages called for by our discussion earlier is bound to be achieved on a particularly simple structure for the strategy. In fact, all the preceding results imply that (D) is a sequential strategy, *i.e.* a sequence of operations of identical character and independent of each other. Let us remark that a problem of the dynamics of organization (cooperative) has been solved by methods which have been developed for the solution of problems of the competitive environment. It is obvious that this methodological possibility is entirely due to the fact that there are only two players with a quite complicated coalition among them.

Macroscopic description again—roles of parameters R and B

Our model of language is fully analogous to the perfect gas of thermodynamics. The only difference is technical and comes from the relative magnitude of possible measuring apparatus and the measured datum. A thermometer is so much larger than the individual molecules that the averaging process which is at the core of the 'measure' of temperature is not as apparent as it is in the 'calculation' of the state function such as B or H, or the state variables such as R or B.

This last is the analogue of the inverse of the temperature as introduced in statistical thermodynamics and the minimization of losses in criterion (d) of p. 496 is entirely analogous to the minimization of free energy to obtain the stable state in thermodynamics. The interpretation is of course opposite: in thermodynamics, the state of equilibrium is the 'worst' one, here it is the 'best'.

As a consequence $\theta = 1/B$ will be called the informational temperature of the text. B can be measured directly on a graph of p_n versus n. However, it is not so for R, the potential number of different words. When the graph of p_n is given, and the canonical law satisfied, one must estimate R by going through the value of some directly measurable number, such as

$$P^{-1}(B, R) = \sum_1^R (n+m)^{-B} \text{ or } H(B, R) = -\sum_1^R p_n \log p_n$$

If $B \leqslant 1$, R must be kept as a second fundamental parameter and can be determined. But this happens only in very exceptional cases, which are in a sense pathological. Thus in modern Hebrew it seems that the H exceeds the $\log R/2$ which is imposed by the Bible.

It turns out that in all the usual cases $B > 1$, and the greater B, the lesser the apparent 'wealth of the vocabulatory'. This makes the estimation of R practically impossible, the finite sums $\sum_1^R$ of a converging series being so close to $\sum_1^\infty$ that the difference is of the order of experimental error. Therefore

R must be taken as 'undetermined', and 'indifferent'. It is not a real characteristic of a text, unless one takes the unrealistic case of a text of infinite length, where *all* words must come out at least once. Therefore B is usually the only important parameter of a text. For B greater but close to 1, H is inversely proportional to $(B - 1)$.

In conclusion, the attempts of some teachers to estimate the 'wealth' of a vocabulary by a 'potential number of words' are futile. Such a notion of 'wealth' is estimated by $1/B$.

However, this has no intuitive connection and one may wish to replace it by some 'apparent value' of R. For example, one could choose the value which would lead to the observed value of H if B were 1, that is Q, such that

$$\log_2 Q/2 = H, \text{ or } Q = 2^{2H}$$

or the square of the number of equiprobable words which would have given the observed value of H.

ACKNOWLEDGEMENTS

The author wishes to express his deep gratitude to Professor G. A. BOUTRY for his support of this, and other fundamental researches, and for permission to publish this paper; and also to express his particular appreciation for the help that D. GABOR extended in the preparation of the paper, and which went much beyond the usual inspection, often amounting to redrafting.

Some of the results had been given in *C.R. Acad. Sci., Paris*, 232 (1951) 1638, 2003. This is a summary of parts of "Contribution à la Théorie Mathématique des Jeux de Communication", *Publ. Inst. Statist, Univ. Paris.*, 2 (1953) 1.

REFERENCES

1 DE SAUSSURE, F., *Cours de Linguistique Générale*, 4th ed., Paris, 1949; (a posthumous collation of lecture notes)
2 ZIPF, G. K., *Human Behavior and the Principle of Least Effort*, (1949)
3 VON NEUMANN, J., *Mathematische Grundlagen der Quantenmechanik*, section V.1 Berlin, 1932
4 SZILARD, L., *Z. Phys.*, 53 (1929) 840
5 GUGGENHEIM, E. A., *Thermodynamics*, (1949)
6 BOREL, E., *Calcul des Probabilités*, Amsterdam, 1921
7 VON NEUMANN, and MORGENSTERN, O., *Theory of Games and Economic Behaviour*, Princeton, 1944
8 WALD, A., *Statistical Decision Functions*, New York, 1950

APPENDIX

Note added after reading the paper by W. H. HUGGINS, p. 363. In other cases, the principle of matching remaining the same, convenience and fruitfulness require the direct method, *i.e.* coding as the unknown. For example, to study the functioning of the ear, Huggins assumes a certain structure of noise and tries to decode it the best way. Strictly speaking, he shows the compatibility of a message and a coding, but for that, he has also to go up on A. COMTE's scales, from mechanics to physiology.

Huggins's theory and ours account for the extreme links of the chain that goes from noise to speech through man. They orient the chain in the same direction. One may hint that the best direction of progress in between will remain the one suitable at both ends of the chain, that is, that Comte's classification will be confirmed by the methodology of the various fields to which the theory of information may be applied.

DISCUSSION

V. BELEVITCH: It has been mentioned that the average length of an English word is 4·5 symbols; as there are 26 symbols in the alphabet, the information content per symbol, neglecting the redundancy, is $\log_2 26 = 4{\cdot}7$, which comes approximately to the same numerical value. The following is an attempt to give a simple theoretical explanation of this coincidence. When the n bits of information contained in an average word are distributed into m symbols of n/m bits/symbol, the process of word recognition is a double selection requiring n/m memory organs to count in the alphabet and m organs to select the various symbols of a word, so that the minimum number of organs is obtained by minimizing $n/m + m$. This is a sum of two terms having a constant product and the minimum occurs for $n/m = m$, which is a known rule for maximum economy of selection equipment in a telephone exchange. The discussion hereabove is made in terms of symbols of an alphabet, but might be repeated in terms of phonemes.

Another theoretical explanation is suggested for the division of the vocabulary into words having an average information content of 20 bits, by analogy with the programme memory of a digital computer. If the capacity of N bits of such a memory is devoted to the storage of M elementary orders (or words) of N/M bits/word, it is required for the complete flexibility of the programme (conditional transfer to subsequences within the programme, *etc.*) that each order may call any other. This is satisfied if the capacity N/M of a word is just sufficient to contain the $\log_2 M$ binary digits necessary to encode the address of any of the M orders. The condition thus yields the transcendental equation $N/M = \log_2 M$ to determine M. If this is true for language, one should be able to deduce, from the known information content per word $N/M = 20$, the number of words in a language by $M = 2^{N/M} = 2^{20} = 10^6$. This is in excess by a factor 5 or so over the number of words in a large dictionary, but gives a right order of magnitude taking into account inflected forms.

A. S. C. Ross: M. Mandelbrot states that 'It is hoped that this work will provide the beginning of a mathematical approach in Linguistics'. His paper is thus of paramount interest to students of Linguistics. It is, indeed, important that there should be liaison between specialists in communication theory and philologists. The gap between the two subjects is very wide, especially in matters of technique and one wonders what philologists are going to make of remarks such as (p. 498), 'Our model of language is fully analogous to the perfect gas of thermodynamics'.

I feel bound to attack M. Mandelbrot in a central point, *viz.* his application of his 'canonical law' to words;

$$p_n = P(n + m)^{-B}$$

All statements of this kind really imply that the occurrence of a word at a given point in a text is a matter of chance and this is what philologists and students of literature will deny. If an English writer has to express the idea 'teapot'—and whether he has to or not is not in the least a matter of chance—the probability of his using the word 'teapot' is unity, and the probability of his using the word 'kettle' is zero.

In M. Mandelbrot's paper, as in communication theory generally, the mathematical concepts are clearly defined. I am afraid that the same cannot be said of the linguistic concepts used by him and other contributors. Thus he makes use of the term 'word'

as if it were possible—or valuable—to define this concept; thus, already, in his Summary, he states that 'language is a message intentionally produced in order to be decoded word-by-word'. Many schools of linguistic scholarship would reject such a view. I myself have always held* that it is neither possible, nor necessary, or even desirable, to define 'word'. And, if we attempt to apply our European ideas of 'word' to other languages, we certainly get into great difficulties. I have suggested (*loc. cit.*) that the fundamental linguistic unit—the unit on which all theory of language must ultimately be based—is something quite other than the word; it is, in my view, a minimal differential unit and is undoubtedly what M. Mandelbrot (p. 490) calls a 'coding dichotomy'. That is to say, the fundamental linguistic units are the least possible differences between pairs of utterances (either 'petty phonematic' as between 'The cat is on the mat' and 'The rat is on the mat', or 'grand phonematic'† as between 'Romulus founded Rome' (plain statement) and 'Romulus founded Rome?' (incredulous question, meaning 'I had thought it was Remus').

M. Mandelbrot states that 'the actual direction of evolution (*sc.* of language) is, in fact, towards fuller and fuller utilization of places'. We are, in fact, completely without evidence as to the existence of any 'direction of evolution' in language, and it is axiomatic that we shall remain so. Many philologists would deny that a 'direction of evolution' could be theoretically possible; thus I myself take the view that a language develops in what is essentially a purely random manner‡. The last part of M. Mandelbrot's sentence is rather obscure, but I seem to sense here the presence of a widespread popular fallacy, that fallacy which may perhaps most concisely be characterized by making assumptions such that English has, in some way, 'progressed' from the inflected state of Anglo-Saxon to its modern almost un-inflected state; or, that there is some essential linguistic difference between a 'primitive' language (such as an Australian one) and a 'civilized' one (such as English). There is, of course, not the faintest support for either of these views.

In his footnote on p. 489, M. Mandelbrot calls attention to de Saussure's famous likening of language to chess. But I think that M. Mandelbrot has misunderstood de Saussure here. The whole point of the parallel‖ lies in his fusing of the two aspects of linguistics thereby, the synchronic (a language at one moment of time) and the diachronic (a language at several moments of time), as he himself calls them§. In a game of chess in play, each piece on the board is in a relation with every other piece on the board, and it is so by reason of moves made in the past. So, too, in a language at a given moment of time, each element of the language is in a relation with every other element of the language and it is so because of the past history—because of the historical philology—of the language. Now in his footnote (as throughout almost the whole of his paper) M. Mandelbrot is clearly—and very naturally—envisaging language from a purely synchronic point of view. But to mention de Saussure's parallel without fusing the synchronic and the diachronic is to lose its whole point. Moreover, I would venture to suggest‡‡ that de Saussure's famous chess-parallel may be more 'neat' than profound.

At the end of his paper, M. Mandelbrot refers to the 'attempts of the linguists to estimate the "wealth" of a vocabulary by the number of words'. I must point out that such estimating, however performed, is no concern of philology or linguistics; it would seem to belong entirely in the domain of statistics.‖‖

* Ross, A. S. C., *Acta Linguistica*, 4 (1944) 101.
† Ross, A. S. C., *Tables for Old English Sound-changes*, pp. 5–6, Cambridge, 1951.
‡ "Sound-change and Indeterminism", *Nature, Lond.*, 129 (1932) 760.
‖ *Op. cit.* (by Mandelbrot), p. 125 ff.
§ In England 'synchronic linguistics' is often called 'descriptive linguistics' and 'diachronic linguistics' is often called 'comparative philology'.
‡‡ See *English and Germanic Studies*, iv, 1–12.
‖‖ With regard to 'number of words' I may refer to my remarks *Actes du Sixième Congrès International des Linguistes* (Paris, 1948), p. 442, and *English and Germanic Studies*, Vol. 4, p. 10.

B. MANDELBROT in reply: I am grateful for Mr. Belevitch's comments as they may provide the starting point for new developments of my theory.

In reply to Prof. Ross, I would say in the first instance that my paper was written primarily for communication engineers and that not every phrase was meant to be useful to philologists. Language, being a canonical message, is really the main point of the paper. Methodologically speaking, statements such as his: 'It is not valuable to define words', cannot be proved by any number of *a priori* arguments, however strong and interesting they may be, but they are fully disproved by the display of a single property, which is satisfied by words, and possibly by other, but not all, elements of speech. I tried to display such a property. I do not 'apply the canonical law to words', but find that Zipf's data on fully inflected words are accurately represented by canonical statistics (and Yule's data on units of the dictionary are not). This disproves Prof. Ross's statement. As to the 'fundamental linguistic units being the least possible differences between pairs of utterances' this is a logical consequence of the fact that two is the least integer greater than one.

How the canonical property is to be 'explained'? My explanation is an example of physical induction: to declare 'true' a theory which leads to correct results. As to Prof. Ross's counter-example of the teapot, it seems to imply that there is always a one-to-one correspondence between de Saussure's 'significants' and 'signifies' and, therefore, that this distinction is pointless. This is true in his example, but in few others: there are many ways of expressing a single idea. In fact, no allusion is made in my paper to any semantic concept, such as 'meaning of the message' and it appears that the correspondence between words and ideas is just loose enough to make the widest variation in meaning compatible with a universal canonical structure.

Again such statements as the last phrase cannot be proved, but are disproved by the actual successful use of this chess-parallel in a limited but specific form.

His other points have been covered, I think, by amendments to the original draft of the paper.

37

SEMANTIC INFORMATION*

YEHOSHUA BAR-HILLEL
Research Laboratory of Electronics, Massachusetts Institute of Technology
and
RUDOLF CARNAP
University of Chicago, U.S.A.

SEMANTIC INFORMATION AND ITS AMOUNTS

THEORY of information, as practised nowadays, is not interested in the content of the symbols whose information it measures. The measures, as defined, for instance, by WIENER and SHANNON, have nothing to do with what these symbols symbolize, but only with the frequency of their occurrence. The probabilities which occur in the definientia of the definitions of the various concepts in information theory are just these frequencies, absolute or relative, sometimes perhaps estimates of these frequencies.

This deliberate restriction of the scope of information theory was of great heuristic value and enabled the theory to reach important results in a short time. Unfortunately, however, it often turned out that impatient scientists in various fields applied the terminology and the theorems of statistical information theory to fields in which the term 'information' was used, presystematically, in a semantic sense, *i.e.* one involving contents or designata of symbols, or even in a pragmatic sense, *i.e.* one involving the users of these symbols. Important as the clarification of the function of the term 'information' in these senses may be, and there can hardly be a doubt as to this importance, 'information', as defined in present information theory, is not a suitable explicatum for these presystematic concepts and any transfer of the properties of this explanation to the fields in which these concepts are of importance may at best have some heuristic stimulating value but at worst be absolutely misleading.

In the following, the outlines of a 'Theory of Semantic Information' will be presented. The contents of the symbols will be decisively involved in the definition of the basic concepts of this theory and an application of the concepts and of the theorems concerning them to fields involving semantics thereby warranted. But precaution will still have to be taken not to apply prematurely these concepts and theorems to fields such as psychology and other social sciences, in which users of symbols play an essential role. It is expected, however, that the semantic concept of information will serve as a better approximation for some future explication of a psychological concept of information, than the statistical concept of present day theory.

We shall attempt to show that the fundamental concepts of the theory of

* This paper will appear in *Brit. J. Phil. Sci.*, Aug. (1953).