Reprinted from SCIENTIFIC PSYCHOLOGY
Edited by Benjamin B. Wolman
© 1965 by Basic Books Publishing Co., Inc.

29

BENOIT MANDELBROT

INFORMATION THEORY AND PSYCHOLINGUISTICS

THIS CHAPTER WILL TELL the story of the brief encounter between two important streams of thought. All the elements of a long and successful association seemed to be present, and the encounter was indeed most exciting. But it was not durable, and the two protagonists have gone their own separate ways, leaving a few fruits that may enjoy some permanency. This chapter will describe a theory of word frequencies thus born out of contact between information theory and psycholinguistics.

The stage having been set, let us introduce the actors.

The great distinction of information theory, born fully armed in 1948 from two celebrated papers by Claude Shannon, was the novel manner in which it recombined two of the oldest and most fundamental ideas of formalized science: the concept of algorithm and the concept of chance. These two ideas have of course never been far from each other, starting with games of pure chance in the seventeenth and eighteenth centuries and continuing through statistical mechanics and quantum theory, around 1900 and 1925, respectively. Therefore, Shannon's theory was in no way the beginning of something entirely new, and now—as time goes by—it begins to fit very well within probability theory as another interesting chapter. However, before 1948, there were only a few scattered attempts to formalize the fact that human discourse is both something highly structured and something highly unpredictable. Shannon's work stressed the possibility of describing its structure by the algorithms of coding and of describing its unpredictability by a systematic exploitation of Markov's pure-chance model of Pushkin's novel *Eugene Onegin;* By reinterpreting the physical concept of entropy as a measure of "quantity of information," Shannon also showed a way of describing the degree of structure of a system by its degree of disorder, itself measured by a function of the probabilities of various events. Of course, it was never claimed that "quantity of information" exhausted the loose verbal idea of "information"; many experts soon found that the best way of presenting Shannon's theory was to follow that author in *not* giving top billing to "quantity of information." It is true, also, that "information theory" was meant to be applicable not only to human discourse, but to every type of message; but the actual techniques of the theory have turned out to be most inconvenient, except for messages similar to discourse.

The importance of Shannon's contribution has always been obvious. But, fifteen years later, it is difficult and somewhat puzzling to think of one's reactions and of those of one's peers, when the original papers appeared, a part of the backlog of novelties accumulated during the war and readied for publication during the three years from 1945 to 1948. The supply and the demand for new things being then at their highest, the ink was not yet dry when enthusiasts started going around promising prompt rewards from the application of information theory to any problem one could think of. Conservatives balked, some saying that it would be too easy to be true, and others adding that one should really not expect that much from a mathematical panacea of which professional mathematicians were taking such an indifferent view. These quarrels have now died, since—as is often the case—most predictions were rather poor. Indeed, application of information theory turned out not to be easy, and it did not solve everything; but the honest toil which it generated has—in our opinion—served as a most effective wedge through which mathematics was helped to conquer ever new fields of application. As to professional mathematicians, they sought to compensate, by great activity, their lateness in getting involved in these problems. To sum up, information theory has indeed become quite well established as a chapter of the calculus of probability. In the meantime, the limitations of its practical usefulness in its field of origin have also become fairly clear.

Let us now turn to the second actor of our story. The term "psycholinguistics" fits him well. But we are not so sure about "linguistics" taken alone, since our story belongs to that margin of problems which were of no concern to the linguists of yesteryear, and of which nobody could know today whether those who will study them tomorrow will choose to be called linguists. What, indeed, is the definition of linguistics? The answer to this question has lately lost the clarity which it seemed once to enjoy. The poet, the philologist, the philosopher, and their traditional associates have ceased to be the only ones concerned with the structure of human discourse. Increasing numbers of mathematicians and of engineers have joined them, attracted by technological problems, formerly ignored or considered trivial, suddenly become so important that they elude traditional barriers between fields, even those which merely contained intellectual curiosity. All this, however, is not enough to make linguistics into what has been called an "interdisciplinary field." In a healthy relation between sciences and technologies, the latter must not determine the limits between the former. Techniques such as applied mathematics may well remain forever at a crossroads between sciences, but the definitions of sciences must be more intrinsic. But, whenever any field enters a period of rapid flux, its definition necessarily becomes vague. Hence, we can neither tell what "linguistics" is really about today nor predict what science or what sciences will be born of the gestations which we are now witnessing. Therefore, we should not venture to prescribe what the word "linguistics" should really mean.

In particular, no one can predict the most important meeting ground between information theory and psycholinguistics, namely, an astonishing statistical law, the discovery of which is associated with the names of Jean-Baptiste Estoup and of George Kingsley Zipf.

Let us again begin with a bit of history. Among the technological problems concerning natural discourse, the most ancient are without question those of cryptography and of stenography or telegraphy. All these problems have now become parts of information theory, but one may question whether they are parts of linguistics. We wish to show that they ought indeed to be considered as such. For that, let us recall the purposes which the cryptographer and the stenographer set for their special transformations of the usual phonic and graphic signs—which are of course quite "arbitrary" in de Saussure's sense. The cryptographer wishes to achieve a code as devoid as possible of any kind of structure that may be used by his adversary to break the secrecy of his message. As to the stenographer and telegrapher, their common aim is to achieve a code in which encoding is as rapid as possible. We must investigate these two kinds of code somewhat more closely.

First of all, we shall neglect the technological constraints which are obviously present in both cases. To that end, let us grant for the moment that the encoding and decoding machines may be as complicated as the designer may wish, and that the memory of the human links—using the common sense of the word "memory"—is unbounded. Under those ideal circumstances, it is obvious that any improvement of our understanding of the structure of language and of discourse will bring a possibility of improvement of the performance of the cryptographer or stenographer. For example, a knowledge of the rules of grammar will show that a given phrase will never be encountered in grammatically correct discourse; thus, if his employer were to speak only grammatical English, a stenographer would not need any special set of signs to designate the incorrect sentences. Similarly, a knowledge of the statistics of discourse will suggest that the "clichés" be represented by special short signs; in this way, the stenogram will be shortened and—since deciphering is very much helped by clichés—the code will be strengthened. That is, the ideal cryptographer and stenographer should make the utmost use of any available linguistic information. Conversely, the empirical findings of language engineers should widen our knowledge of language and of discourse. (Moreover, this interplay between theory and practice should be widened to include the role of another group of language engineers: those concerned with helping translation from one language to another with the help of automatic dictionary look-up. However, the scope of the present work excludes this question and any other topic of the field of "mechanical translation.")

Let us indulge at this stage in a brief epistemological aside, pointing out that one could hardly have expected the empirical facts most important to language engineers to be those that most interest traditional linguists. As a matter of fact, the following is likely to happen and has indeed been observed: using the margin of error allowed by empirical observation, gramarians and language engineers may very well envision their common object of study in such divergent ways that their final theories are, strictly speaking, logically incompatible. The discovery of such occurrences has had a rather traumatic effect on some linguists, whose previous experience did not include logical deductions sufficiently long to ever lead to logical contradiction. Practitioners of supposedly "hard" fields, such as physics, are on the contrary very familiar with the fact that a total description of a given reality may require the use of several logically incompatible theories. There is also good reason for the psychologist to envy the physicist's luck, in his attempts to explain complicated facts on the basis of very simple assumptions: after all, practitioners of the "hard science" of physics have succeeded—for centuries—in being excused for using arguments about the imaginary realm of atoms; no one could check the properties assumed for these entities, so that the philosophers' strictures against such "non-operational" procedures were paid little heed to.[1]

Feeling sure that nobody will take quite seriously this aside, let us resume our examination of the cryptographer and the stenographer. It is clear that they work under such obvious practical constraints that they could seldom take advantage of the possibilities of making use of liguistics, which we have described. Conversely, the professional literature of those fields is hardly known to outsiders, so that their discoveries are, hardly well known. There are a few exceptions; from the viewpoint of information theory, the most important are the frequency properties based either on "average samples of discourse" from mixtures of various sources, or on samples from well-determined authors. The striking fact here is the difference in simplicity between the two most commonly studied "articulations," namely, that of the letter or phoneme, and that of the word.

The frequencies of single letters, and of "*n*-grams" made up of *n* suc-

[1] This is a good example of what the physicist Eugene P. Wigner refers to as "the unreasonable effectiveness of mathematics in the natural sciences." How embarrassing, by contrast, is the situation encountered in the social sciences: any set of asssumptions —however fruitful, reasonable, and useful it may be—is suspectible of some kind of experimental verification, which—however rough it may be—is very likely to show the inapplicability of these assumptions.

cessive letters, have been taken into account by cryptographers and telegraphists since earliest times. Decades before the justification provided by information theory, Samuel Morse knew that he should use the shortest combinations of dots and dashes to designate the most frequent letters; cryptographers dealt with letter frequencies centuries before information theory. Moreover, recurrent attempts have been made to relate the frequency of a phoneme to some measure of its articulatory "difficulty." But all this has not taken us very far. Indeed it seems that, from the viewpoint of statistical model-making, isolated letters are too small to be intrinsic units: little is implied about longer bits of discourse by knowing the frequencies of its contributing letters. As to *n*-grams, they are more useful, but linguistically very artificial.

Words are better bets for the theoretician: although their exact linguistic standing is not without problems, there is no question that perception of discourse is based on units of the approximate length of a word, and there is good practical reason for beginning the study of those units by examining the words defined as the sequences of letters between two successive space-symbols.

Such a study was performed quite early; according to our sources, the first worker in this area was a stenographer of the French Parliament, Jean-Baptiste Estoup. His work seems to have been motivated by a politico-technical dispute concerning the respective advantages of several systems of French stenography and—very commendably—he sought scientific fact to support the design which he favored. Similar investigations were repeated, this time for the sake of healthy intellectual curiosity, by many other investigators. But theirs were isolated and brief efforts in comparison with the lifetime of toil devoted to the question of word frequencies by George Kingsley Zipf, the author of several books that combine fact and folly in an unusually intimate fashion. To describe the findings of these various authors, take a long sample of discourse from a given individual, and rank all the words that occur in this sample in the order of decreasing frequency. The word given rank 1 is the most frequently observed sequence of letters contained between two successive space-signs; in English, it is usually "the," but for some subjects it is "I." The word given rank 2 is that which becomes the most frequent when the word of rank 1 has been put aside. The word of rank 3 is the most frequent when one excepts those of ranks 1 and 2, and so on. Let us designate by the symbol $W\ (r)$ the word that occupies the rank number r in this special ordering. We should note that, when one gets down to rare words, one finds many that occur only once or twice in the given sample. The order of rare words is indeterminate but also indifferent; that is, these words can be ranked arbitrarily. Using the definitions given above, the empirical findings can be expressed as follows:

In the first approximation, the ratio $i(r,k)/k$, which is the relative number of repetitions of the word $W(r)$ in the sample of length k, is inversely proportional to 10 times r:

$$i(r,k)/k = 1/(10r)$$

This is supposed to hold for every writer regardless of the language that he uses. The usual way of checking relationships of this form is to use logarithmic paper, in which the abscissa is the logarithm of r and the ordinate is the logarithm of $i(r,k)$. The first-approximation law of word frequencies is expressed by stating that the graph of log $[i(r,k)]$ as a function of log r is a straight line of slope -1, that is, parallel to the second bisectrix of the coordinate axes, as shown by the solid line of Figure 29-1.

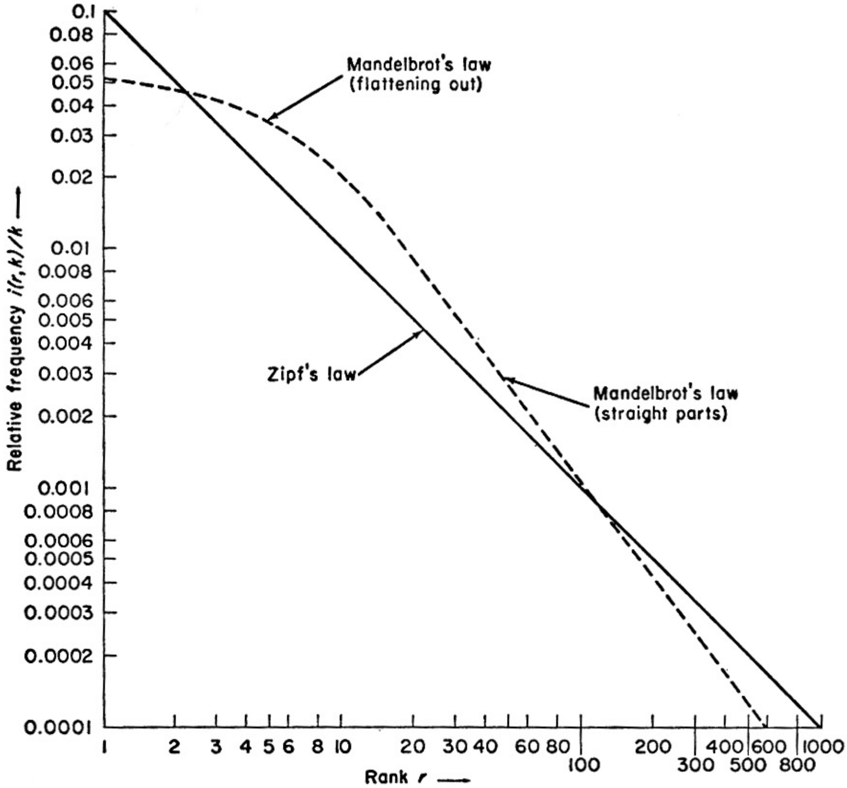


Figure 29-1.

In the second approximation, one finds that most empirical graphs differ markedly from a straight line of slope -1. *First of all,* the first few words seem not to follow the law at all. As a matter of fact, in languages such as French, the definition of "word" is unclear in the case of abbreviated forms such as *l'*, which are among the most frequent, so that "the" distribution of the most frequent words is ill determined. One finds, moreover, that the bulk of the graphs of log $[i(r,k)]$ are not parallel to the second bisectrix, and that their slope depends upon what is loosely referred to as the "wealth of vocabulary" of the subject. Such a dependence is hardly unexpected. To sum up, data can be analytically represented by the following formula (the reader may prefer to skip this formula; we hope that he will be able to follow the sequel anyway):

$$i(r,k) = Pk(r+V)^{-B}$$

The rank, r, has been defined. As to P, V, and B, they are "parameters": that is, they are fixed for a given subject, but are different for different subjects; they do not characterize a language, although it may well be that different languages favor different ranges of values for the parameters. The easiest to measure is parameter B, which is the absolute value of the slope of the logarithmic graph of log $[i(r,k)]$ as a function of log r (excluding the most frequent words). In the first approximation, $B = 1$ and $V = 0$.

Social science statistics being what it is, the foregoing second approximation is among the best-established results of that field. One still finds some authors, however, who say that it is either false, absurd, or self-evident; one author even says that it is partly false, partly absurd, and partly self-evident.

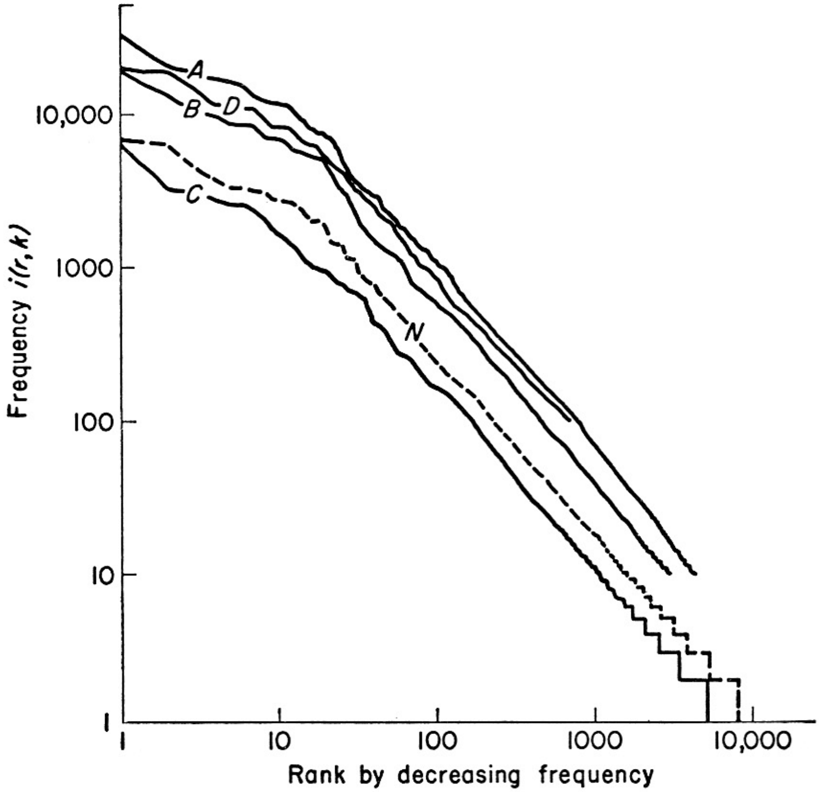
But actually there is nothing obvious in the law of word frequencies: Of course, by their very definition, the quantities r and $i(r,k)$ vary in inverse *directions,* but this is not the same as to say that they vary in inverse *proportion,* as suggested by the first approximation. As a matter of fact, distributions of the form above are closely related to probability laws having an infinite first moment; without needing to know the exact meaning of this technical term, the reader may be assured that the occurrence of such laws was felt until very recently to be practically never observed in practice, rather than to be obvious.

Note also that the double logarithmic graphs have been so misleading, at times, that they have come to be distrusted at all times; actually, their validity depends entirely upon the value of the coefficient B, and they are very safe in the range of values of B which are observed for word frequencies.

The final criticism of our distribution is that it is a priori absurd to imagine that any general law could be applicable to word frequencies. But there is actually no such conflict between the facts and most people's "intuition" of the properties of discourse. Let us discuss this in detail.

To fix our ideas, consider the word "coffee." It is usually observed when the subject wishes to express some idea, without reference to any desire to attribute a given frequency to the sequence of letters "c-o-f-f-e-e." But, first of all, the probabilistic conception of the structure of discourse determines the relative frequency of a word only on the average. To say that "coffee" occurs exactly once every 1,770 words implies belief in a nonstochastic structure. This is very similar to well-known facts concerning the tossing of coins or of dice: if the ace appears exactly once for each play in a long succession of plays (each made up of 6 tosses), one may be sure that the game is unfair. In the ideal model of random tosses, the observed frequencies of the ace must vary (or fluctuate) around the ideal value (1/6). Similarly, the observed frequencies of words should be expected to fluctuate. Naturally, as seen by the emitter, the fluctuations of word frequencies are not generated by "pure chance" but rather are associated with what he "has to say"; but, as seen by the receiver, the occurrences of the words are (hopefully) at least partly unpredictable, and "pure chance" is nothing but a model of unpredictability. One may very well find that this model is inadequate for certain refined purposes, but one can certainly not say that it is a priori absurdly in conflict with any kind of intuition.

Let us proceed. The law $i(r,k) = Pk(r + V)^{-B}$ does not limit itself to saying that every word has a well-defined frequency; it also proposes a



See page 558 for explanation of this part of the figure.

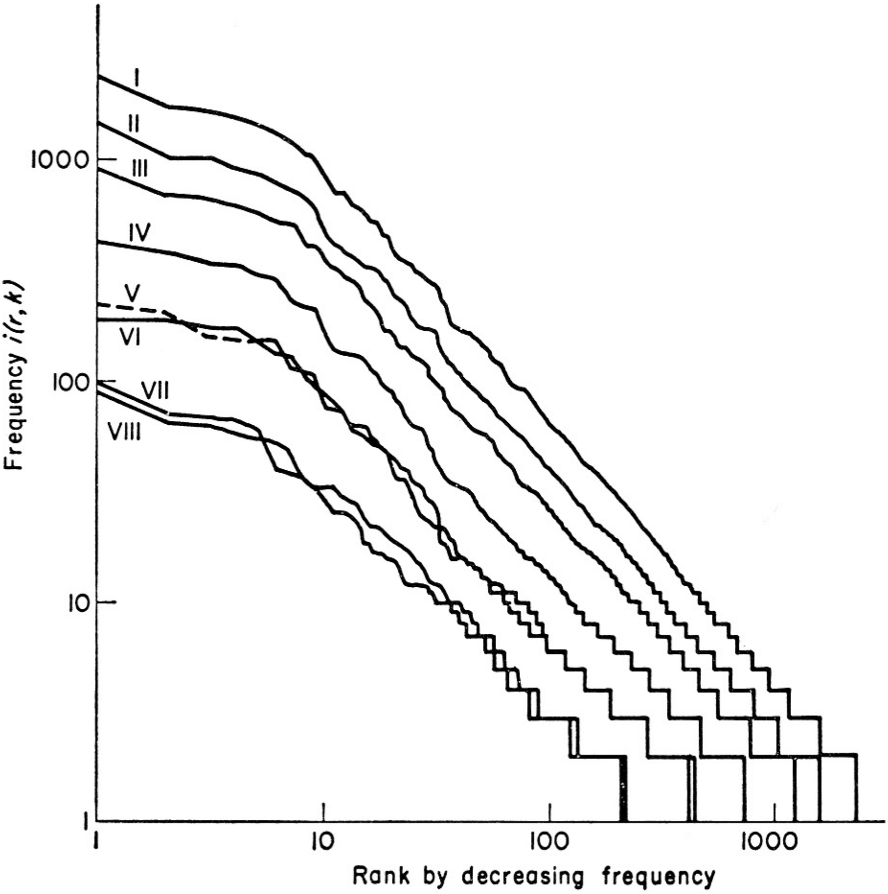


Figure 29-2. Some examples of plots of the frequency of a word (vertically in the logarithmic scale) versus its rank (horizontally in the logarithmic scale). *First: A, B, C,* and *D* are samples from German writers, N, from a Norwegian writer; *second:* samples of increasing length from the writings of the same English-speaking individual, who happened to be a schizophrenic.

relation between the frequencies of different words. This is not absurd either. First of all, the law says nothing of the frequencies of "coffee," "tea," or "chocolate" as such. Its prediction is rather the following: Paul tells me that he has established that "coffee" has rank 177 in his vocabulary; whereas Peter found that it has rank 315 in his. In the first approximation, this suggests to me that Paul uses the word "coffee" once every 1,770th word on the average, and that Peter uses it every 3,150th word. That is, the law of word frequencies relates to the entire system of Peter's or Paul's words, taken as one whole. Moreover, "system of words" is in no way synonymous with "system of ideas." The correspondence between the two is sufficiently "arbitrary" to allow one system of words to represent many systems of ideas, and conversely.

Having hopefully convinced the milder skeptics of the reasonableness of the existence of a law of word frequencies, it remains to "explain" why this law takes the form which we have described above. It happens that this expression is, formally, one of the simplest encountered by statisticians; but most professionals neglected it until recently, so that almost all the explanations (beginning with Zipf's "least effort" arguments) used very primitive tools, or handled the mathematics incorrectly. Hence—so far as we know—the only acceptable models are the numerous variants of an idea which I suggested in 1951 and which was developed in various directions by several writers. These variants are fully equivalent mathematically, but they appeal to such different intuitions that the strongest critics of one may be the strongest partisans of another; certain variants were in fact rediscovered by constructive critics! (I have long since given up trying to determine which is best.) It is useful to note here that a problem of prediction is called "well posed" in physics if slight changes in initial conditions do not change the result more than slightly. Similarly, our "explanatory" setup may be called well posed, since changes in assumptions do not lead to disproportionate changes in the results. (This outlook is not as obvious as it may seem: one incorrect model of the law of word frequencies considers as an asset the fact that, by changing its assumptions slightly, one could also obtain any one of several other laws which share none of the deeper properties of the distribution actually observed.) Let us now examine a few of the variants of our model.[2]

Two of the variants are readily interpreted in terms of the problems of cryptography and telegraphy mentioned earlier. We can prove this: suppose that the cryptographer or the telegraphist are forced to use a given coding alphabet and that they must encode word by word, the code of each word being followed by a special symbol that plays the same role as the "space" of ordinary spelling. Then, it is shown by information theory that the same rule of coding is best for both secrecy and economy (because any symbols that could be spared for purposes of economy might also provide the clue for deciphering). Let us now examine the constraint imposed by the use of word-by-word coding. This requirement may be expected to yield decreased economy and decreased strength, as compared to coding procedures that cut across words and use longer units. However, in the case of actually observed word statistics, and in this case alone, no loss whatsoever is entailed by the need to use word delimitations and to follow

[2] For the details of the various derivations, the reader is referred to Mandelbrot, 1961. Less complete account but probably better-written is the account given in Miller and Chomsky, 1963. For further developments addressed to the philosopher of science, see Mandelbrot, 1957.

the words by "space" symbols. That is, we may say that, insofar as the use of a space is intrinsic to the concept of a word, the only case where the word is a natural segment of discourse is when its statistics follow the law which they happen precisely to follow.

The form of these criteria is very familiar in physics, which is quite inclined to characterize the observed facts with the help of such statements as "the principle of least action," "the principle of largest entropy," "the principle of smallest entropy production," and the like. Of course, the initial idea of those principles was borrowed from introspective criteria of optimal human behavior, and, after a long detour through physics, they returned to social science under the guise of such things as Zipf's "principle of least effort." As a matter of fact, I acknowledge that it is the title of Zipf's book that triggered my first derivation of the law of word frequencies, a model in which "telegraphic optimality" was characterized by maximizing Shannon's "quantity of information" under certain constraints. Unfortunately, it seems that many readers were greatly confused by the inevitable association of such a model with Zipf's peculiar idea of a perfect world and by the inevitable association of Shannon's "quantity of information" with the many aspects of the word "information" (which Shannon never remotely claimed to cover). It turned out, therefore, to be more politic to stress the optimality of observed word statistics with respect to cryptography, an activity less readily associated with the idea that the works of man are perfect.

A third variant of our basic model also deserves brief mention. It is clear indeed that, as time goes by, the structure of the system of word frequencies slowly changes, both for individuals and for the averages relative to groups. Concerning the nature of this change, I made certain hypotheses that are simple and would have seemed reasonable in the framework of the "unreasonable effectiveness of mathematics in natural science" to which we have referred. Since, as mentioned, the model thus obtained is only a reinterpretation of the two criteria of "optimality," it leads again to the empirically observed law. Again, there is little chance of a universe generated by haphazard diachronic chance being considered "perfect" in any real way.

As we near the end of our tale, we realize that we have not done justice to psychology as such. Let us therefore end by commenting on a curious and controversial aspect of some of our models. Independently of the specific variant that one may prefer, our models ultimately rest upon the decomposition of words into more elementary units. For example, in the last analysis, the cryptographic or telegraphic optimality implies the representation of words with the help of such special signs as Morse dots and dashes. As a matter of fact, this decomposition into parts is the key to the success of our models since, using appropriate new forms of (roughly) the "law of large numbers," we could show that the same law of word frequencies should correspond to a wide range of microscopic structures; this is an application to linguistics of a method of "macro-model" making in the absence of "micro-data," which is one of the most powerful tools of the physicist (a different tool bearing the same name is greatly used in economics). This being granted, and even though word frequencies are surely independent of the technical details of Samuel Morse's contraption, one could have expected them to be linked to phoneme frequencies and—in the first approximation—to letter frequencies. For example, most of our variants require a concept of "the cost" of a word, and it would be tempting indeed to identify the cost of a word with the number of letters which it contains. Unfortunately, this is impossible, and the best that one can do is to look for the cost within the recoding of discourse by the higher nervous system of the receiver of the message and perhaps even of its emitter.

Even though hardly anything is known about those stages, it is likely that this recoding uses certain "units" shorter than the phrase or the "idea" and longer than the phoneme or the letter. It is natural to try to determine whether these units are the words; if so, our models would refer to the mutual "adaptation" between the codes of those words and their frequencies. Unfortunately, the only simple tests of this hypothesis are the tachistoscopic measurements of the time required in order to identify words. These tests are favorable to our hypothesis, but the recourse to higher brain functions may still seem to be a dodge. Let us hope, however, that the situation is no worse than that of statistical physics in the heyday of the energetists: those who dared oppose the philosophers of science spoke freely of the effects of the shape of the unobservable atoms; the more cautious scholars preferred to work with the so-called "phenomenological" method, for which theories that predict well the results of possible experiments need not be "explained" any further. As it happens, I personally am rather in favor of phenomenology, these days, so that we shall not continue our tale any further.

REFERENCES

Estoup, Jean-Baptiste. *Les gammes sténographiques.* Paris: privately printed for the Institut Sténographique, 1916.

Mandelbrot, Benoit. Adaptation d'un message à la ligne de transmission. *Comptes Rendus des Séances Hebdomadaires de l'Académie des Sciences de Paris,* 1951, **232,** 1638-1640.

Mandelbrot, Benoit. Linguistique statistique macroscopique. In L. Apostel, B. Mandelbrot, and A. Morf, *Logique, langage et théorie de l'information.* Paris: Presses Univ. de France, 1957, Pp. 1-80.

Mandelbrot, Benoit. On the theory of word frequencies and on related Markovian models of discourse. In Roman Jakobson (Ed.), *Structure of language and its mathematical aspects, proceedings of symposia on applied mathematics*. Vol. 1. Providence, R. I.: American Mathematical Society, 1961.

Miller, George A., & Chomsky, A. Noam. Finitary models of language users. In R. R. Bush, E. Galanter and R. D. Luce (Eds.), *Handbook of mathematical psychology*. Vol. 2. New York: Wiley, 1963.

Shannon, Claude. A mathematical theory of communication, *Bell System Technical J.*, 1948, **28,** 379-423, 623-656.

Zipf, George Kingsley. *Human behavior and the principle of least effort.* Reading, Mass.: Addison-Wesley, 1949.