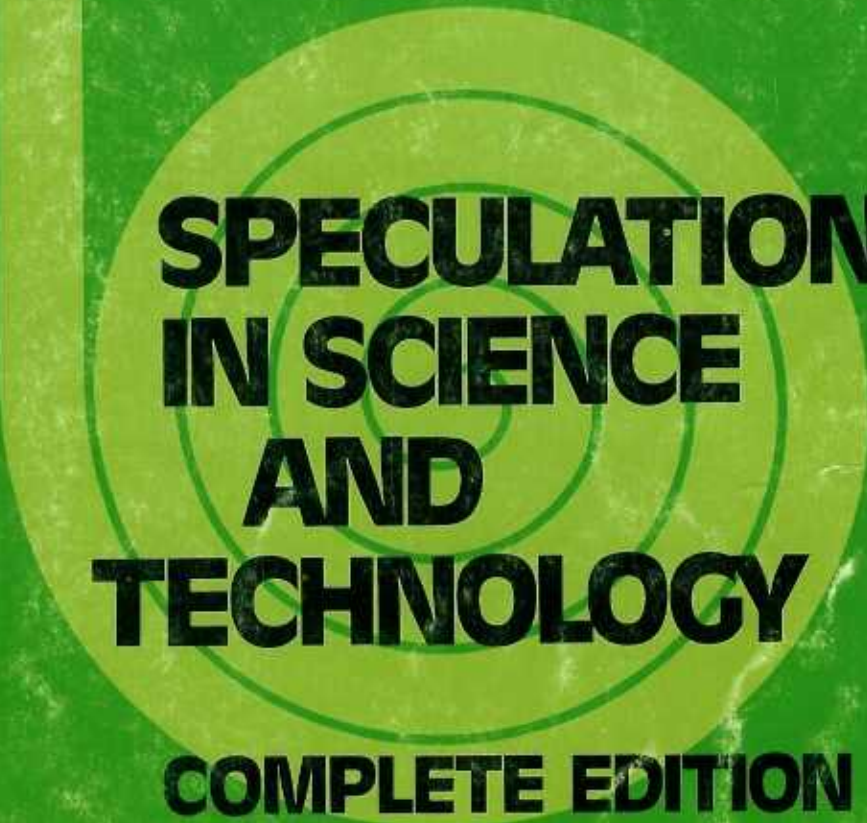



Published five times per annum

ISSN 0155-7785

Volume 6 Number 2

June 1983



# **SPECULATIONS IN SCIENCE AND TECHNOLOGY**

## **COMPLETE EDITION**

Editor and Founder:  
William M. Honig



Elsevier Sequoia - Lausanne

## THE GENESIS MODEL — PART II: FREQUENCY DISTRIBUTIONS OF ELEMENTS IN SELFORGANIZED SYSTEMS

PETER WINIWARTER

*Le Bordalier Institute, F-41270 Droue, France.*

*Received: 7 September 1982*

### Abstract

A quantitative measurement of the complexity  $C$  of a selforganized system implies the determination of the frequencies of the element types constituting the system. The analysis of empirical data concerning a large variety of selforganized systems — such as cosmic systems, biological systems, economic systems and linguistic systems — suggests a general law. *The second law of genesis:* The rank-ordered frequency distribution of element-types of any selforganized system at any level of hierarchical organization can be approximated by a mathematical distribution of the form  $p_n = A(n + m)^{-B}$ .

### 1. INTRODUCTION

In part I of the genesis model<sup>(1)</sup> we have introduced a quantitative measure for the complexity of a selforganized system of matter  $C = IR$ . The postulated first law of genesis  $\Delta C \geq 0$  implies, that for a system with constant energy redundancy  $R$  the average selective information per energy state increases ( $\Delta I \geq 0$ ) and the system is at equilibrium if a maximum value  $I_{\max}$  corresponding to the maximum value of the entropy  $S_{\max}$  is reached.

Let us now consider the logical counterpart, that is consider a system with a constant value of  $I$ . Applying the first law of genesis  $\Delta C \geq 0$  implies  $\Delta R \geq 0$ . Such the energy redundancy of a system with constant information should increase and the system is at equilibrium, if a maximum value  $R_{\max}$  is reached.

In analogy to Boltzmann's approach it would be interesting to calculate the theoretical frequency distribution of elements  $p_i$  corresponding to  $R_{\max}$  and a given value of  $I$ , using the method of Lagrange multipliers. This would imply that the functional relationship of the quantity to be maximized  $R$  and the different relative frequencies  $p_i$  be known explicitly. In our definition of the energy redundancy

$$R = \left[ \left( \sum_{i=1}^M n_i e_{i0} - E_0 \right) / \sum_{i=1}^M n_i e_{i0} \right] \times 100 [\%]$$

where  $n_i$  the absolute frequency of the element-type with the energy at rest or rest-mass  $e_{i0}$ , and  $E_0$  the energy at rest or rest-mass of the total system, the dependence of  $E_0$  on the distribution  $n_i$  or  $p_i = n_i/N$  is not known explicitly.

Leaving aside this theoretical hurdle, we have chosen a path demanding "less effort" by limiting ourselves to the analysis of empirical data.

## 2. ZIPF'S ANALYSIS OF HUMAN LANGUAGE

"Human Behaviour and the Principle of Least Effort" is the title of a study of Zipf<sup>(2)</sup> published in 1949. We ignore whether Shannon's and Weaver's "Mathematical Theory of Communication"<sup>(3)</sup> published in book form in the same year have influenced the works of Zipf, but we suppose that — as so often in the history of science — the discovery of Zipf's empirical law and Shannon's quantitative definition of selective information have evolved independently. A summary of Zipf's tedious but remarkable study relevant to the ideas proposed in this paper is given by Cherry<sup>(4)</sup>;

"Figure 1 shows curve A, the result of statistical analysis made upon James Joyce's *Ulysses*; the volume contains about a quarter of a million word tokens with a vocabulary of nearly 30,000 word-types. (*Token* is the name of every individual word that actually appears in printed text; *type* refers to the entries of a vocabulary list or dictionary of the text.) This curve A results from plotting the frequencies of the various word-types against their rank-order. (In statistical studies, if a number of elements are listed in decreasing order of their frequencies of occurrence  $f_1, f_2 \dots f_n \dots$  then they are said to be *rank-ordered* in frequency. The suffixes 1, 2 . . . n . . . may be regarded as units on a linear scale of rank-order.) Several aspects of this curve are remarkable. Naturally, this curve must slope downward from left to right, but we have no right whatsoever to assume that any part of it would be at all smooth — let alone straight. It might well descend from left to right in a series of irregular jumps; again, rather than approach a straight line, it might take the form of a dotted curve as C or D.

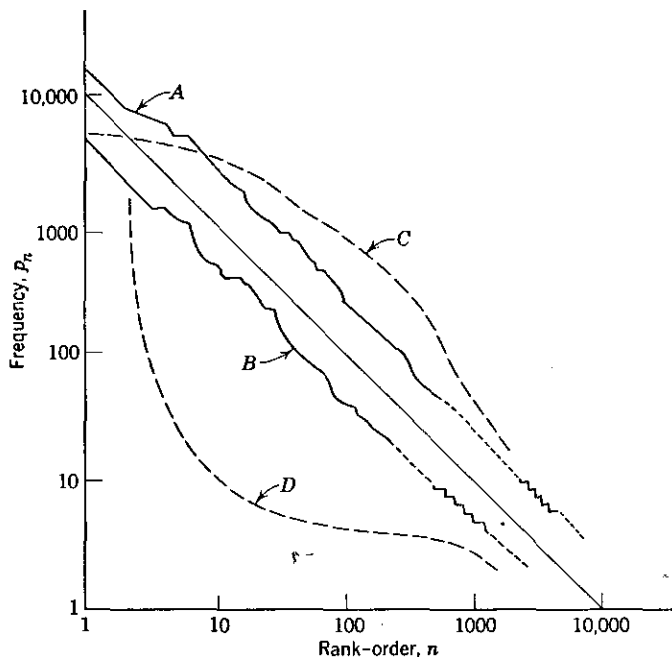


Figure 1. The rank-frequency distribution of words: (A) James Joyce's "Ulysses"; (B) American newspaper English; (C) and (D) hypothetical (after Zipf).

Such a linear law is derived from empirical data; if the source of data be changed markedly, it may be felt that the change would be reflected in the form of the law. But Zipf takes some different data, corresponding to samples of American newspapers, and plots them as in curve B. Considering the divergent natures of these sources of language, the two curves A and B are surprisingly similar. Zipf reinforces his evidence for the existence of a definite "law" by amassing similar data from widely different languages of the world (for example, see Figure 2 below), and from texts covering a thousand years of history. Not only words but other segments of text have been studied in such a statistical manner; phonemes, syllables, morphemes — and even Chinese characters, and the babbling of babies."

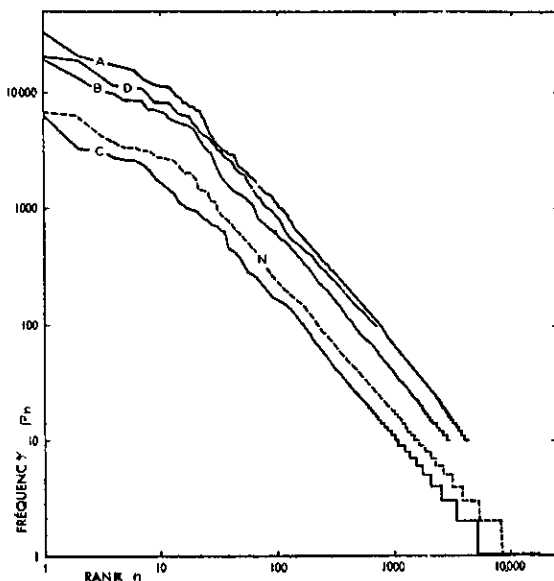


Figure 2. The rank-frequency distribution of words: A—C Norwegian; N German (after Zipf).

### 3. MANDELBROT'S EXPLICATION OF ZIPF'S LAW

In a detailed study Mandelbrot<sup>(5)</sup> shows that Zipf's empirical law

$$p_n = A n^{-B} \tag{1}$$

where A and B are constants and  $B \approx 1$  or a law, which approximates the empirical data even better

$$p_n = A (n + m)^{-B} \tag{2}$$

where

$$A = 1 / \sum_n (n + m)^{-B}$$

B = const., and m is a parameter, can be derived as the result of an optimization making the following assumptions:

- (a) The number of letters  $M_1$  and the number of potential words  $M_w$  are constant.
- (b) The average selective information per word of the considered system is given in advance and constant.

$$I = - \sum_n p_n \log_2 p_n = \text{const.}$$

- (c) The quantity to be optimized/minimized is the "average cost per word" of the system:

$$\bar{c} = \sum_n p_n c_n$$

where  $c_n$  is a hypothetical "cost" of the word-type occurring with the frequency  $p_n$  measured in arbitrary units of entiers.

Assuming that the "cost" of a word depends on the "costs" of its constituting letters only, and making various assumptions about the "costs" assigned to letters (all letters of equal cost, any cost assigned to the various letters of the alphabet — or the cost of any letter in a word depending upon the preceding letter) Mandelbrot shows, that in all cases the cost of the  $n$ -th group of letters ordered according to increasing "cost" can be generally expressed in the form:

$$c_n \cong c_0 + \log_{M_1} (n + m) \quad (3)$$

The method of Langrange multipliers yields for the minimum of the "average cost per word"  $\bar{c}$  given constant  $M_1$ ,  $M_w$  and  $I$  a distribution of the form:

$$p_n = A e^{-B' c_n} \quad (4)$$

Replacing  $c_n$  in (4) by expression (3) yields a distribution which can be approximated by (2). Such in bilogarithmic coordinates the expression  $-\log p_n = -\log A + B \log(n + m)$  describes most of Zipf's empirical data in a satisfying way.

Mandelbrot's approach is very close to the general problem posed in section 1, the only difference being that we wanted to determine the frequency distribution  $p_n$  yielding a maximum of the energy redundancy  $R$  of a system with given constant  $I$ , while Mandelbrot determines the frequency distribution yielding a minimum of the average hypothetical "cost" per word for a system with given constant  $I$ .

Assuming that Mandelbrot's hypothetical "average cost per word"  $\bar{c}$  is linked to our definition of the energy redundancy  $R$  of the system through the simple relationship  $\bar{c} \propto (1 - R)$ , a minimization of  $\bar{c}$  corresponds to a maximization of  $R$ . Expression (2) can therefore be considered as the theoretical solution describing the frequency distribution of a system of words with a given constant value  $I$  at equilibrium, that is corresponding to the maximum value of  $R = R_{\text{max}}$ .

#### 4. THE SECOND LAW OF GENESIS

Mandelbrot's analysis limits itself to systems constituted of words; each word being an ensemble of letters separated from other words by space.

We conjecture that the arguments developed above can be extended to any selforganized system at any level of organization and postulate a general law.

*The second law of genesis:* The rank-ordered frequency distribution of subsystems/elements of any selforganized system at any hierarchical level of organization can be approximated by a mathematical distribution of the form

$$p_n = A (n + m)^{-B} \quad (2)$$

where  $A = 1 / \sum_n (n + m)^{-B}$ ,  $B = \text{const.}$ ,  $n$  the rank-order and  $m$  a parameter influencing the distribution for small  $n$  only.

In the following we will put forward some evidence in favour of this hypothesis.

## 5. FREQUENCY DISTRIBUTION OF CHEMICAL ELEMENTS IN THE UNIVERSE

In the analysis of section 3, we considered language as a selforganized system constituted of elements (words) separated by space. Defining mutually exclusive types of elements (word-types) in terms of the number of components and their sequence at the next lower hierarchical level (letters of an alphabet) we determined the rank-ordered frequency distribution of element-types, which can generally be approximated by a mathematical law of form (2).

In analogy we consider the universe as a selforganized system constituted of elements (atomic isotopes) separated by space. Defining mutually exclusive types of elements (chemical elements) in terms of the number of components and their sequence at the next lower hierarchical level (electrons, protons) we determined the rank-ordered frequency distribution of chemical element-types. Figure 3 (on the following page) based on the data compiled by J.P. Meyer and A.G.W. Cameron as quoted by H. Reeves<sup>(6)</sup> indicates that this distribution can be described by a law of type (2). Except for  $n < 7$  (influence of the parameter  $m$ ), the empirical data can surprisingly well be approximated by a straight line.

## 6. FREQUENCY DISTRIBUTIONS IN BIOLOGICAL SYSTEMS

(i) *A single species constituted of ensembles of individuals.* Considering a local animal population belonging to a single species of parasites and their hosts as a selforganized system constituted of elements (ensembles of individuals on single hosts) separated by space, we define mutually exclusive types of elements in terms of the number of individuals in an ensemble, that is by the number of parasites on a single host.

Figure 4 (on the following pages) shows an example of a rank-frequency distribution of parasite ensembles based on the data quoted by C.B. Williams<sup>(7)</sup>.

(ii) *A single genus constituted of species.* Considering a local insect population belonging to a single genus only as a selforganized system constituted of elements (species), we define mutually exclusive types of species in terms of the number of individuals constituting each species.

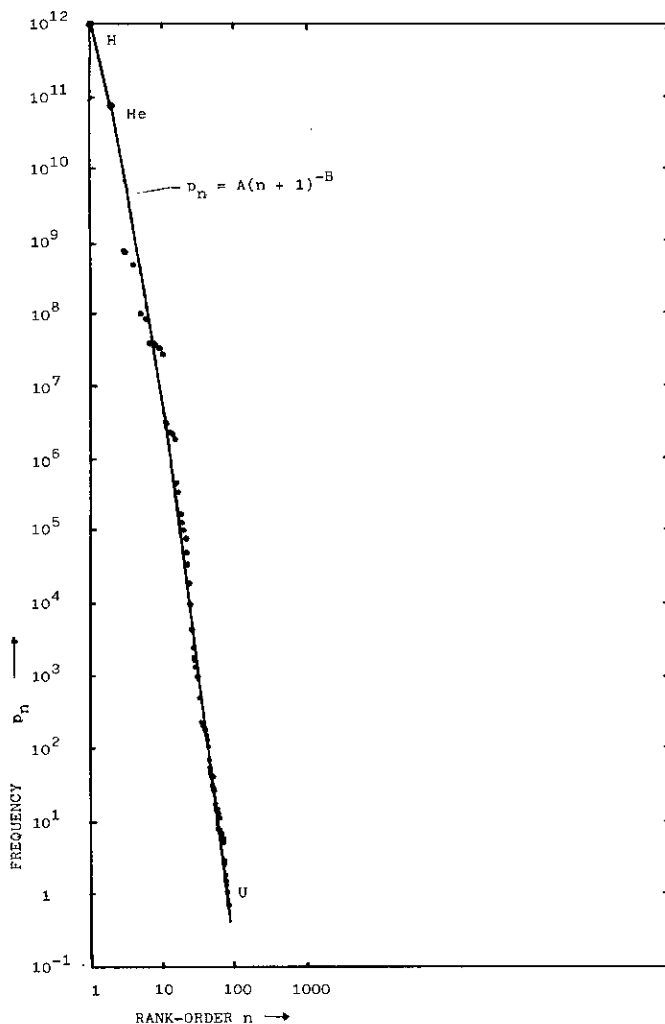


Figure 3. The rank-frequency distribution of chemical elements in the universe.

Figure 5 (opposite) shows a rank-frequency distribution of species-types based on the data of C.B. Williams<sup>(7)</sup>, which are the result of daily random samples of the genus *Macrolepidoptera* caught in a light trap at Rothamsted during four successive years (patience again!). Altogether there were 15,609 individuals caught, representing 240 species.

(iii) *A single family constituted of genera.* In analogy to (ii) one can climb up the hierarchy and consider the animal population belonging to a single family as a selforganized system constituted of elements (genera) and define mutually exclusive genus-types in terms of the number of species constituting each genus. Figure 6 (on the following pages) shows a rank-frequency distribution of genus-types based on the classification of the insect family *Mantidae* by W.F. Kirkby (in 1910) as quoted by C.B. Williams<sup>(7)</sup>.

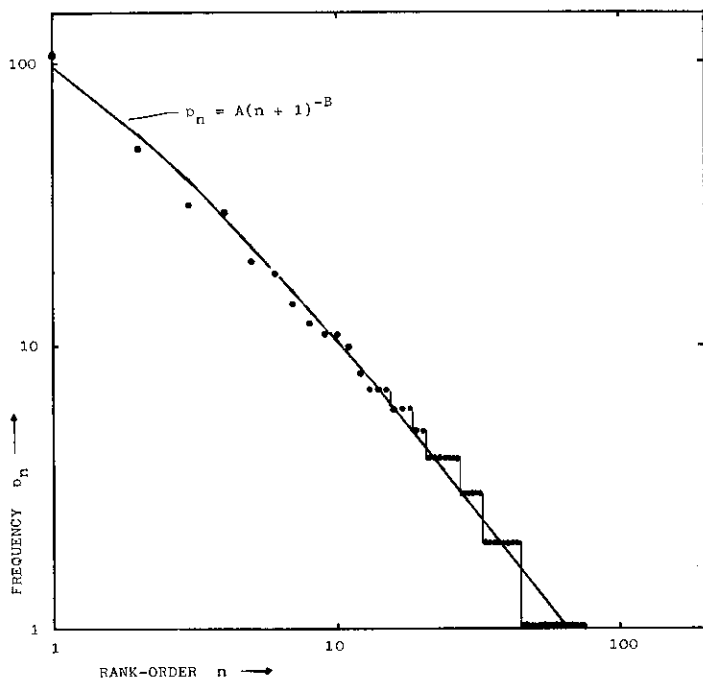


Figure 4. Rank-frequency distribution of ensembles of individuals belonging to one biological species.

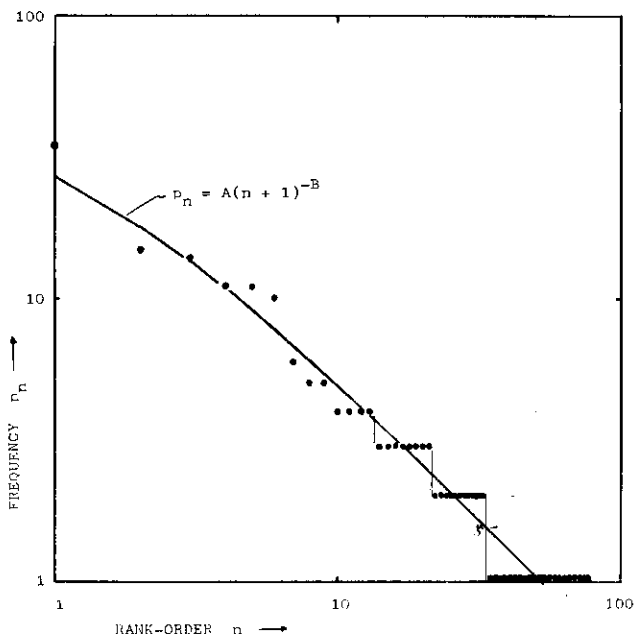


Figure 5. Rank-frequency distribution of species belonging to one biological genus.



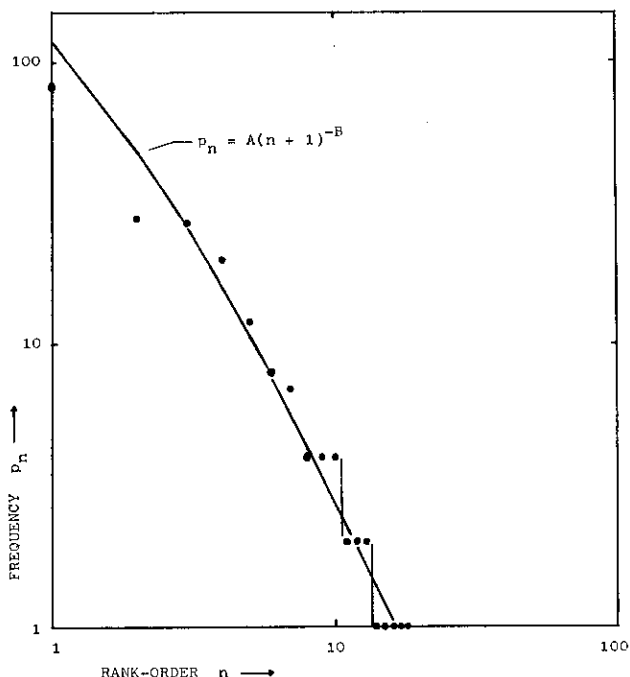


Figure 6. Rank-frequency distribution of genera belonging to one biological family.

## 7. FREQUENCY DISTRIBUTIONS IN ECONOMIC SYSTEMS

Considering a capitalist economy as a selforganized system constituted of elements (monetary units representing services or goods, like words representing abstract notions or physical objects) we define mutually exclusive types of elements; each element type being characterized by the enterprise producing it.

Figure 7 (opposite) shows the rank-frequency distribution of monetary units produced by the 1000 greatest French business enterprises based on the data of the year 1980 compiled by Dun & Bradstreet as published by C. Barjonet<sup>(8)</sup>.

Except for  $n < 7$  (influence of the parameter  $m$ ) the 1000 empirical data points can be approximated by a straight line in double-logarithmic coordinates with an accuracy rarely found in experimental sciences.

## 8. CONCLUSION OF PART II

Despite the small sample of systems analyzed, taking into consideration their divergent nature, we can say that the empirical data speak in favour or — given the rather poor statistics for biological systems — do not obviously contradict the highly speculative second law of genesis.

A further preliminary survey concerning self-organized systems of matter has revealed the following result:

- the rank-frequency distributions of chemical elements in the universe, of masses in the solar system (sun + planets), of masses in the Saturn system (planet + satellites), of chemical elements in the earth shell, of chemical elements in sea-water and of chemical elements in the human body can all be fairly well described by distributions of form (2); and what seems even more surprising — the slopes of the straight lines in bilogarithmic coordinates, corresponding to the constants B, are very similar and could reveal identical within error limits.
- on the other hand the slopes of the straight lines describing the rank-frequency distribution of ensembles of individuals of a biological species in an eco-system, of words in a linguistic system and of monetary units in an economic system are also very similar and could be described by a single constant.

In view of these preliminary results, the hypothesis put forward in Part I, that one and the same *quantitative* principle governs the process of evolution at all levels ( $\Delta C \geq 0$ ) seems to us deserving of further inquiry.

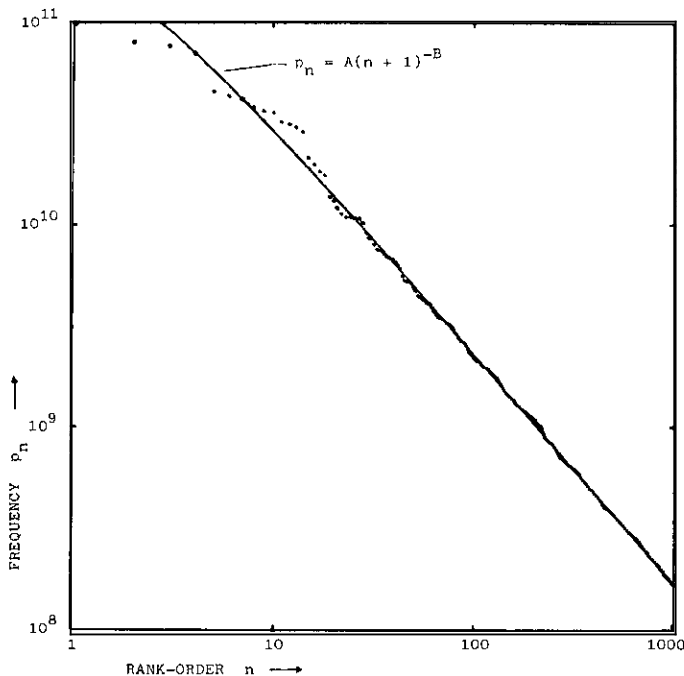


Figure 7. Rank-frequency distribution of monetary units in a capitalist economy.

## 9. OUTLOOK

Part III will present the actual genesis model, which has led to the development of our concept of a quantitative measure of complexity. Introducing the notion of abstract automata — a mathematical concept initially derived from the study of sequential switching circuits — the process of evolution could be described as the selforganisation of one abstract automaton following one and the same algorithm throughout its organization into more and more complex subsystems or abstract sub-automata.

## Acknowledgements

I would like to express my gratitude to all the scientists whose patient compilation work has made the above quick survey possible.

## References

1. Winiwarter, P., *Spec. Sci. Tech.*, 6, 11 (1983).
2. Zipf, G.K., *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Cambridge, Mass. (1949).
3. Shannon, C.E. and Weaver, W., *The Mathematical Theory of Communication*, University of Illinois Press, Urbana (1949).
4. Cherry, L., *On Human Communication*, M.I.T. Press, Cambridge, Mass. (1966).
5. Mandelbrot, B., *Contribution a la theorie mathematique des jeux de communication*, Thesis Sc.Math, Paris (1952). No. 3393 published in *Extr. des Publications de l'institut de Statistique de l'universite de Paris*, 2, fasc. 1 and 2, Paris (1953).
6. Reeves, H., *Patience dans l'azur, L'evolution cosmique*, Editions du Seuil, Paris (1981).
7. Barjonet, C., *L'Expansion*, 6, 109-309 (1981).