PUBLICATIONS DE LA SOCIETE SCIENTIFIQUE

EXTRAIT DU CATALOGUE

ANNALES DE LA SOCIETE SCIENTIFIQUE DE BRUXELLES,

t. I à XLVI, 1875 à 1926. Chaque vol. in-8° de 400 à 600 pages . . F **150,00**

ANNALES DE LA SOCIETE SCIENTIFIQUE DE BRUXELLES,

Série A (sc. mathématiques), t. XLVII à LVI (1927 à 1936) . . . . . F **70,00**
Série B (sc. physiques et naturelles) » » » F **70,00**
Série C (sc. médicales) (1927 et 1928) F **100,00** — (1929 à 1933) (1931 à 1936) F **40,00**
Série D (sc. économ. et techniques) (1927 à 1929) F **20,00** — (1930) (1931 à 1936) F **60,00**
Série I (sc. mathématiques et physiques), tt. LVII à LXVII (1937 à 1953) F **70,00**
t. LXVIII à LXXII (1954 à 1958) F **100,00**
t. LXXIII (1959) F **150,00**
Série II (sc. naturelles et médicales), tt. LVII à LX (1937 à 1940, 46) F **70,00**
Série III (sc. économiques), tt. LVII à LX (1937 à 1940, 46) . . . . . F **100,00**

REVUE DES QUESTIONS SCIENTIFIQUES,

t. I à XCII (1877 à 1927). Les deux volumes annuels . . . . . F **200,00**
Le fascicule trimestriel . . . . . F **60,00**
t. XCIII à CXVI (1928 à 1939). Les deux volumes annuels . . . . F **200,00**
Le fascicule . . . . . F **40,00**
t. CXVII (1940 et 1946), tt. CXVIII à CXXX (1947 à 1959). Le volume F **200,00**
Le fascicule . . . . . F **60,00**

TABLES ANALYTIQUES DES ANNALES,

t. I à XXV (1875 à 1901) . . . . . F **20,00**
t. XXVI à XLVI (1902 à 1926) . . . . . F **40,00**

TABLES ANALYTIQUES DE LA REVUE,

t. I à L (1877 à 1901). . . . . F **20,00**
t. LI à LXXX (1902 à 1921) . . . . . F **20,00**
t. LXXXI à CX (1922 à 1936) . . . . . F **30,00**

MONOGRAPHIES DE SCIENCES NATURELLES

I — **B. Tougarinoff.** Les réactions organiques dans l'analyse qualitative minérale (cations). — Un vol. in-8° de 107 pages (1930) : en Belgique, F **24,00**; autres pays : F **30,00.**

II — **V. Schaffers.** Le paratonnerre et ses progrès récents. Un vol. in-8° de 90 pages (1931) : en Belgique, F **24,00**; autres pays : F **30,00.**

IV — **F. Kaisin et E. de Pierpont.** — Hydrogéologie des Calcaires de la Belgique. Un vol. in-8° de 111 pages, avec 35 fig. et un plan hors texte (1939) : en Belgique, F **24,00**; autres pays, F **30,00.**

MONOGRAPHIES MEDICALES

I — **M. Schillings.** Le rein en fer à cheval. Un vol. in-8° de 104 pages, avec 8 planches hors-texte (1938) : en Belgique, F **70,00**; autres pays, F **90.00.**

III — **P. Van Gehuchten.** La pathologie du système pallido-strié. Un vol. in-8° de 52 pages, avec 8 planches hors-texte (1930) : en Belgique, F **24,00**; autres pays, F **30,00.**

MONOGRAPHIES DES SCIENCES ECONOMIQUES

I — **A. Henry.** La structure technique de l'agriculture belge et ses particularités en Wallonie et en Flandre. Un vol. de 66 pages . . . . . F **20,00**

II — **A. Henry.** Les variations régionales de l'Agriculture en Belgique. Un vol. de 50 pages . . . . . F **10,00**

III — **A. Delpérée.** La réglementation conventionnelle des conditions de travail en Belgique. Un vol. de 200 pages . . . . . F **60,00**

IMPRIMÉ EN BELGIQUE

Annales de la SOCIÉTÉ SCIENTIFIQUE de Bruxelles

Association sans but lucratif

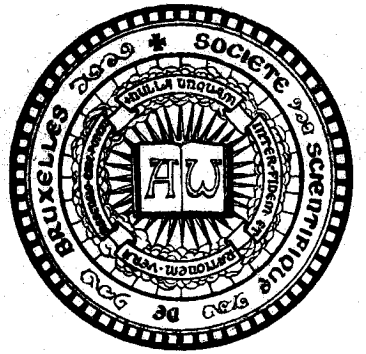
TOME SOIXANTE-TREIZIÈME

SÉRIE I

SCIENCES MATHÉMATIQUES ASTRONOMIQUES ET PHYSIQUES

TROISIÈME ET DERNIER FASCICULE

21 décembre 1959



Publié avec le concours de la Fondation universitaire de Belgique et du Gouvernement

SECRÉTARIAT DE LA SOCIÉTÉ SCIENTIFIQUE
11, RUE DES RÉCOLLETS, 11
LOUVAIN

1959

Publication trimestrielle. Prix de ce fascicule séparé : 70 frs

Annales de la Société Scientifique de Bruxelles, Tome 73, nº III, p. 310-326

Deuxième Section

SCIENCES PHYSIQUES ET CHIMIQUES

On the statistical laws of linguistic distributions

BY

V. BELEVITCH (*)

SUMMARY

The rank-frequency diagrams of statistical linguistics are reinterpreted as distribution curves of the cumulative probability of types in the catalogue versus the probability of tokens in the text. For such distributions, the closure condition $\Sigma p_i = 1$ (which does not hold in general statistics for the independent variable) imposes certain relations between the mean, the variance, the number of elements in the catalogue and the average information content (negative entropy). Sections 2 to 4 are devoted to the mathematics of these relations, especially to their particular forms for truncated normal distributions. First and second order Taylor approximations to an arbitrary distribution law take the form of Zipf's and Mandelbrot's laws respectively. Experimental data lead to accept the truncated normal distribution with $\sigma \cong 2{,}8$ bits as the general law for words. Data on letter and phoneme distributions seem to indicate that the standard deviation has the universal value $\sigma \cong 1{,}4$ bits.

1. RANK-FREQUENCY DIAGRAMS

It is an experimental fact that by counting frequencies of occurrence of various elements (letters, phonemes, words) in homogeneous texts written in a given language, and dividing them by the total number of elements of the same nature in the text, one often obtains relative frequencies that are stable (independent of the length of the text for sufficiently long texts). These relative frequencies define the a priori probabilities of the elements. Having used *texts* to obtain probabilities of various linguistic elements, one constructs *catalogues* of elements of identical nature (i.e. alphabet for letters, lexicon for words, etc.) in which each

(*) Comité d'Etude et d'Exploitation des Calculateurs Electroniques « C.E.C.E. », 67, rue de la Croix de Fer, Brussels, Belgium.

element is listed with its probability, by order of non-increasing probabilities.

In normal statistical practice one defines discrete distributions by specifying the number N_i of elements having a dimension or some other measurable characteristic x_i. If $N = \Sigma N_i$ is the total number of elements, the ratio $f_i = N_i/N$ is the relative frequency, or probability of finding the value x_i for the dimension x. The cumulative probability is defined by

$$\varphi_i = \sum_{k=1}^{i} f_k \tag{1}$$

and distribution curves are obtained by plotting the cumulative probability φ_i versus the dimension x_i.

Most linguistic elements (f.i. letters) have no measurable characteristic (dimension) according to which they could be ordered, except their text probabilities themselves. As there is no point in defining a distribution curve by a text probability versus a text probability, the only alternative is to plot the probability in the catalogue f_i versus the probability in the text p_i. The probability in the catalogue is defined by reference to experiments with an urn containing once each element i, marked with its text probability p_i : the catalogue probability f_i is the probability of drawing from the urn the text probability p_i. For a catalogue of N elements, N_i of which have the text probability p_i, f_i is the ratio N_i/N, and the cumulative probability (1) in the catalogue is

$$\varphi_i = \frac{1}{N} \sum_{k=1}^{i} N_k \tag{2}$$

But the sum ΣN_k in (2), i.e. the number of elements of text probabilities $\leqslant p_i$, is precisely the rank i in the catalogue, since elements are ranged in order of non-increasing frequencies. As a consequence (2) becomes $\varphi_i = i/N$ and gives the *relative rank* in the catalogue. The distribution curves thus defined only differ from the usual rank-frequency diagrams, where the rank of each element in the catalogue is plotted versus its probability in the text, by the factor $1/N$ transforming absolute rank into relative rank. As a conclusion, *rank-frequency diagrams can be interpreted as ordinary distribution curves giving the cumulative probability in the catalogue versus the probability in the text.* This establishes a relation between the paradigmatic and syntagmatic aspects of the language.

Information theory attributes to each element of probability p_i an information measure (negative entropy) $x_i = -\log p_i$, and it is therefore convenient to use logarithmic scales for text probabilities. When a logarithm of base 2 is used in the above definition, the information is measured in *bits*. In the analytical expressions, it is more convenient to use natural logarithms; this is equivalent to adopt $\log_2 e = 1{,}44$ bits as natural unit of information (*binit*).

In general statistics, the range of the independent variable x is unrestricted. In statistical linguistics, x is related to a probability by

$$x = -\log p \tag{3}$$

and is essentially positive. An additional restriction results from the *closure condition*

$$\sum_1^N N_i p_i = 1 \tag{4}$$

which is transformed into

$$\sum N_i e^{-x_i} = 1 \tag{5}$$

or, since the probability in the catalogue was defined as $f_i = N_i/N$, into

$$\sum f_i e^{-x_i} = \frac{1}{N} \tag{6}$$

As a conclusion, *the closure condition determines the absolute number of elements in the catalogue from their relative distribution law*

2. Mean values

When dealing with mean values, one should clearly distinguish between *averages over the text* and *averages over the catalogue*. The mean m and the variance σ^2 as defined by

$$m = \frac{\Sigma N_i x_i}{N}; \quad \sigma^2 = \frac{\Sigma N_i (x_i - m)^2}{N} \tag{7}$$

are averages over the catalogue. In the text, the number N_i of elements of measure x_i must be weighted proportionally to their probability of occurrence p_i, and the average information is

$$h = \frac{\Sigma N_i p_i x_i}{\Sigma N_i p_i}$$

Since the denominator is unity, this gives the mean information

$$h = -\sum N_i p_i \log p_i \tag{8}$$

in the sense of information theory. Finally (8) becomes

$$\frac{h}{N} = \sum f_i x_i e^{-x_i} \tag{9}$$

We now consider the case where the range of x, for which f takes significant values, is sufficiently small, i.e. *narrow distributions*, so that the function e^{-x} can be approximated by the Taylor expansion

$$e^{-x} = e^{-m} [1 - (x-m) + \tfrac{1}{2} (x-m)^2] \tag{10}$$

around the mean m. Condition (6), combined with (7), yields

$$e^{-m} (1 + \sigma^2/2) = 1/N \tag{11}$$

Similarly, in (9), the expansion of $x e^{-x}$ is deduced from (10) by writing $x = (x-m) + m$ and neglecting the third order term. One finds

$$xe^{-x} = me^{-m} \left[1 - \left(1 - \frac{1}{m}\right)(x-m) + \left(\frac{1}{2} - \frac{1}{m}\right)(x-m)^2\right]$$

and (9) becomes

$$\frac{h}{N} = me^{-m} \left[1 + \left(\frac{1}{2} - \frac{1}{m}\right)\sigma^2\right] \tag{12}$$

Since σ^2 is small for a narrow distribution, (11) becomes approximately

$$m = \log N + \tfrac{1}{2} \sigma^2 \tag{13}$$

On the other hand, the ratio of (12) and (11) gives, with the same approximation,

$$h = \log N - \tfrac{1}{2} \sigma^2 \tag{14}$$

or

$$h = m - \sigma^2 \tag{15}$$

If all N elements have the same text probability $1/N$, the information is the constant $\log N$. By comparison, formulae (13) and (14) show that *the effect of a small spread in the probabilities is to reduce the average information in the text* (and this is well known from information theory) *and to increase by the same amount the average over the catalogue*. Furthermore, *the variance of the distribution is* precisely *the difference between both averages*.

It is often convenient to approximate discrete distributions with a large number of elements by continuous distributions. The number of elements of dimension comprised between x and $x + dx$ is $N\, d\varphi(x) = N\, \varphi'(x)\, dx = N\, f(x)\, dx$, where $\varphi(x)$ is the distribution function and $f(x) = \varphi'(x)$ the probability density. A Gaussian, or normal, distribution of mean m and variance σ^2 is defined by the probability density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{16}$$

The range of the normal distribution is $-\infty \leqslant x \leqslant \infty$ and, since x is essentially positive in linguistic applications, the distributions cannot be rigorously normal. In the following we will consider truncated normal distributions where (16) is restricted to some positive interval $x_a \leqslant x \leqslant x_b$, the density assuming the value 0 outside. We will start, however, by the case where the truncation is made at points sufficiently far away from the mean, so that the tails of the distribution are negligible anyway.

For a continuous distribution, conditions (6) and (9) are replaced by

$$\frac{1}{N} = \int_{x_a}^{x_b} f(x) e^{-x}\, dx\,; \quad \frac{h}{N} = \int_{x_a}^{x_b} x e^{-x} f(x)\, dx \tag{17}$$

In these expressions also, we first replace the integration limits by $\pm\infty$, neglecting the truncation. Absolute convergence is still ensured, for the increase of $|x|$ and e^{-x} for $x = -\infty$ is less rapid than the decrease of (16).

For the density (16), the integrals (17) are reduced to the error integral by transforming the exponent according to

$$x - \frac{(x-m)^2}{2\sigma^2} = -\frac{(x+\sigma^2-m)^2}{2\sigma^2} - m - \frac{\sigma^2}{2} \tag{18}$$

By the linear transformation

$$z = \sigma + \frac{x-m}{\sigma} \tag{19}$$

the first equation (17) becomes

$$\frac{1}{N} = \frac{e^{-m+\sigma^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = e^{-m+\sigma^2/2} \tag{20}$$

and is equivalent to (13). The second equation (18) becomes

$$\frac{h}{N} = \frac{e^{-m+\sigma^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (m - \sigma^2 + \sigma z) e^{-z^2/2}\, dz \tag{21}$$

The term in z in the integrand is an odd function and does not contribute to the integral, so that (21) reduces to

$$\frac{h}{N} = (m - \sigma^2) e^{-m+\sigma^2/2} \tag{22}$$

and, by comparison with (20), one obtains (15).

The remaining part of this section is devoted to the derivation of the rigorous formulae replacing (20) and (22) and taking the truncation into account. For the error integral, we will use the notation

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-u^2/2}\, du \tag{23}$$

so that the distribution function corresponding to the density (16), with neglected truncation, is

$$\varphi(x) = \Phi\left(\frac{x-m}{\sigma}\right) \tag{24}$$

If truncation is taken into account, the density cannot be (16), but must be corrected by a factor $a > 1$ in order to normalize to unity the integral in the finite range. Similarly the distribution function is no longer (24) but is of the form

$$\varphi(x) = a\Phi\left(\frac{x-m}{\sigma}\right) - b \tag{25}$$

deduced from the density (16) multiplied by a, with an integration constant denoted $-b$. These constants are determined by the conditions $\varphi(x_a) = 0$, $\varphi(x_b) = 1$, and one obtains

$$a = \frac{1}{\Phi\left(\frac{x_b-m}{\sigma}\right) - \Phi\left(\frac{x_a-m}{\sigma}\right)}\,; \quad b = \frac{\Phi\left(\frac{x_a-m}{\sigma}\right)}{\Phi\left(\frac{x_b-m}{\sigma}\right) - \Phi\left(\frac{x_a-m}{\sigma}\right)} \tag{26}$$

For the truncated distribution, the integrals (17), applied to the density $a\,f(x)$, become

$$1/N = ac\,e^{-m+\sigma^2/2} \tag{27}$$

$$\frac{h}{N} = ae^{-m+\sigma^2/2}\left[c(m-\sigma^2) + \frac{\sigma}{\sqrt{2\pi}}\left(e^{-z_a^2/2} - e^{-z_b^2/2}\right)\right] \tag{28}$$

where z_a and z_b have the values resulting from (19) with $x = x_a$ or x_b, and where c denotes the error integral with the limits z_a and z_b, thus

$$c = \Phi\left(\sigma + \frac{x_b - m}{\sigma}\right) - \Phi\left(\sigma + \frac{x_a - m}{\sigma}\right) \tag{29}$$

In (25), the term b introduces a correction at high probabilities, which is negligible in most applications because $x_a - m$ is negative and equals several times σ. On the contrary, the effect of the truncation at low frequencies is often not negligible, for statistics do not generally extend sufficiently far above the mean. The practical form of (27) is thus deduced from the approximation $x_a = -\infty$, and is

$$N = e^{m-\sigma^2/2}\,\frac{\Phi\left(\dfrac{x_b - m}{\sigma}\right)}{\Phi\left(\sigma + \dfrac{x_b - m}{\sigma}\right)} \tag{30}$$

In a complete statistical count, the lowest frequency corresponds to a single occurrence in the text, thus to probability $p_b = 1/L$, where L is the length of the text. By (3), one has

$$x_b = \log L \tag{31}$$

and (30) gives a relation between the length of the text and the extension of the vocabulary. It is obvious that (30) reduces to (13) for $x_b = \infty$.

4. Approximations to normal distributions

We consider first an arbitrary distribution function $\varphi(x)$ in the neighbourhood of a point x_0. The Taylor expansion is

$$\varphi(x) = \varphi(x_0) + (x - x_0)\,f(x_0)$$

where $f(x)$ is the corresponding probability density. The logarithmic slope of distribution function is approximately

$$\log\frac{\varphi(x)}{\varphi(x_0)} = \log\left[1 + \frac{f(x_0)}{\varphi(x_0)}(x - x_0)\right] \cong (x - x_0)\,\frac{f(x_0)}{\varphi(x_0)} \tag{32}$$

If one considers an element $x = x_i$ corresponding to a text probability $p_i = e^{-x_i}$, its rank i is given by $N\varphi(x_i)$. Introducing the similar notations p_0 and i_0 for the reference point x_0, (32) becomes

$$\log\frac{p_i}{p_0} = -A\log\frac{i}{i_0} \tag{33}$$

with

$$A = \frac{\varphi(x_0)}{f(x_0)} \tag{34}$$

Equation (33) is independent from any assumption on the distribution law, and merely shows that (34) measures the slope of the tangent, at x_0, to the rank-frequency characteristic with logarithmic scales for both coordinates.

Expression (33), or

$$\frac{i}{i_0} = \left(\frac{p_i}{p_0}\right)^{-1/A} \tag{35}$$

is similar to Zipf's law, but with a variable slope. By taking into account second order terms in the Taylor expansion, it is possible to obtain a correction similar to the one introduced in Zipf's law by Mandelbrot, i.e. to arrive at a form

$$\frac{i}{i_0} = s\left(\frac{p_i}{p_0}\right)^{-1/B} - t \tag{36}$$

instead of (35), or, equivalently, to

$$p_i = P(i + \rho)^{-B} \tag{37}$$

with

$$P = p_0(i_0 s)^B;\quad \rho = i_0 t \tag{38}$$

The best values of the parameters are obtained by identifying the second-order Taylor expansion of the right-hand side of (36), i.e.

$$s e^{(x_i - x_0)/B} - t = s - t + s\,\frac{x_i - x_0}{B} + \frac{s(x_i - x_0)^2}{2B^2}$$

with the corresponding expansion of the distribution function

$$\frac{\varphi(x_i)}{\varphi(x_0)} = 1 + \frac{(x_i - x_0) f(x_0)}{\varphi(x_0)} + \frac{(x_i - x_0)^2 f'(x_0)}{2\varphi(x_0)}$$

This gives

$$s = \frac{[f(x_0)]^2}{\varphi(x_0) f'(x_0)}; \quad t = s - 1 \tag{39}$$

$$\mathrm{B} = \frac{f(x_0)}{f'(x_0)} \tag{40}$$

and one has $t > 0$ as long as the curvature of the characteristic at x_0 is positive.

For a normal distribution (truncated or not), (40) becomes

$$\mathrm{B} = \frac{\sigma^2}{m - x_0} \tag{41}$$

Simple expressions for the other parameters are only obtained if some approximations are introduced, and truncation neglected. For the linguistic applications it is of special importance to discuss the behavior of the characteristics at high frequencies, where the statistics are the most reliable. We will thus assume $x_0 \ll m$ and use the asymptotic expansion (1)

$$\Phi(-u) = \frac{e^{-u^2/2}}{u}\left(1 - \frac{1}{u^2} + \dots\right) \tag{42}$$

valid for large positive values of u. When the first term alone of (42) is considered, the value (41) is obtained for the exponent (34) of the Zipf approximation. With two terms in (42), (39) gives

$$s = 1 + \frac{\sigma^2}{(m - x_0)^2}; \quad t = \frac{\sigma^2}{(m - x_0)^2} \tag{43}$$

By (15), (41) becomes

$$\mathrm{B} = \frac{\sigma^2}{\sigma^2 + h - x_0} \tag{44}$$

When the approximating point x_0 is the point of highest frequency x_a, (44) gives $\mathrm{B} < 1$, since the average information h is certainly larger than the minimum information x_a. When x_0 increases

(1) The possibility of deducing Mandelbrot's law with $\mathrm{B} \leqslant 1$ from the symptotic expansion of the error integral was mentioned to the writer by A. OETTINGER, and originated the present investigation.

starting from x_a, B increases and passes through the value 1 for $x_0 = h$.

5. LETTER AND PHONEME DISTRIBUTIONS

Because of the small size of the alphabets, such distributions are relatively irregular, but definite systematic trends can, however, be noticed. On the other hand, thanks to the small size of the alphabets, the statistics are more reliable, and the distributions are well known over their entire ranges. Fig. 1 shows data from a number of languages (2); the lower scale is logarithmic in p, and

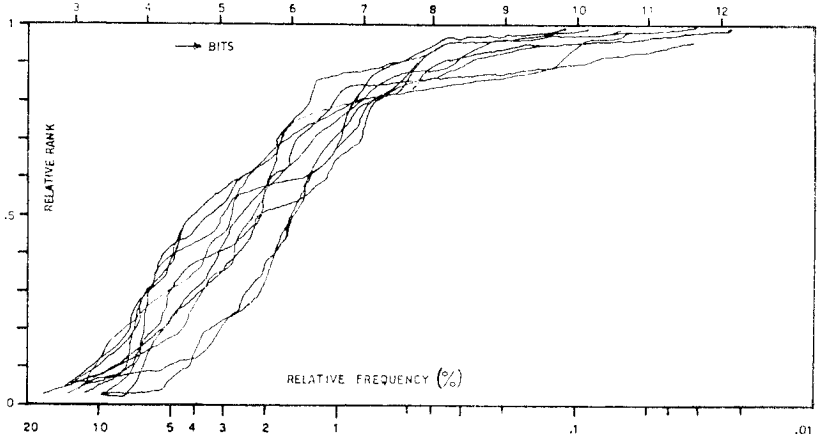


Fig. 1

the upper scale is linear in bits. Its appears immediately that all curves are very similar, and all ranges extend from approximately 2.5 to some 11 bits. All distributions are relatively narrow, and it is therefore expected by (13) that the horizontal shift between the various curves is correlated with the size of the alphabet (varying from $\mathrm{N} = 21$ to 49 in the examples considered). This is indeed the case, as it appears from Fig. 2 where each curve has been shifted by $\log \mathrm{N}$; in other words, the abscissa in fig. 2 is the relative frequency p/p_m with respect to the equiprobable value $p_m = 1/\mathrm{N}$.

(2) Most of these have already been discussed in V. BELEVITCH « Théorie de l'information et statistique linguistique », *Bull. Acad. Roy. Belg.* (Cl. des Sc.) avr. 1956 pp. 419-436.

Fig. 2 clearly shows that all distributions are practically identical and, in particular, have the same variance. Our best estimate of the common value is $\sigma = 1{,}4$ bits. The corresponding normal distribution shifted by $\sigma^2/2$ relatively to the point of abscissa log N in accordance with (13), is the dotted curve of fig. 2.

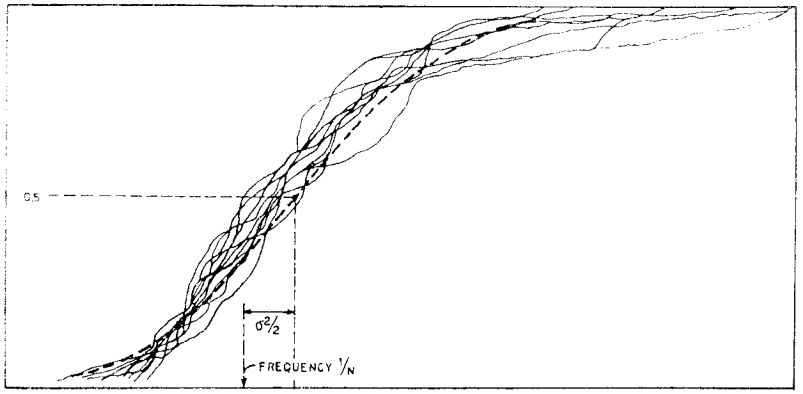


Fig. 2

Formulae (13-15) can be checked on the example of Russian phonemes for which particularly detailed data are available ([3]). The value of log N is $\log_2 42 = 5{,}42$ bits, and the published value of h is 4,78 bits. From (15), one obtains $\sigma = 1{,}33$ bits, and from (13), $m = 6{,}1$ bits; the last value agrees with the median of the distribution curve deduced from the published data.

6. Word distributions

According to Guiraud ([4]), Zipf's formula, and even Mandelbrot's correction, do not agree with most experimental distributions at low frequencies. The truncated lognormal character of the actual distributions seems to have been suspected by several authors ([5]). If truncation is neglected, a lognormal distribution

([3]) E.C. Cherry, M. Halle, R. Jakobson, « Toward the logical description of Languages in their phonemic aspect », *Language*, vol. 29 n° 1, p. 34, March 1953.

([4]) P. Guiraud, *Les caractères statistiques du vocabulaire*, Paris, Press. Univ. 1954.

([5]) See f. i., J. Aitchison, J.A.C. Brown, *The lognormal distribution*, Cambridge 1957; p. 101.

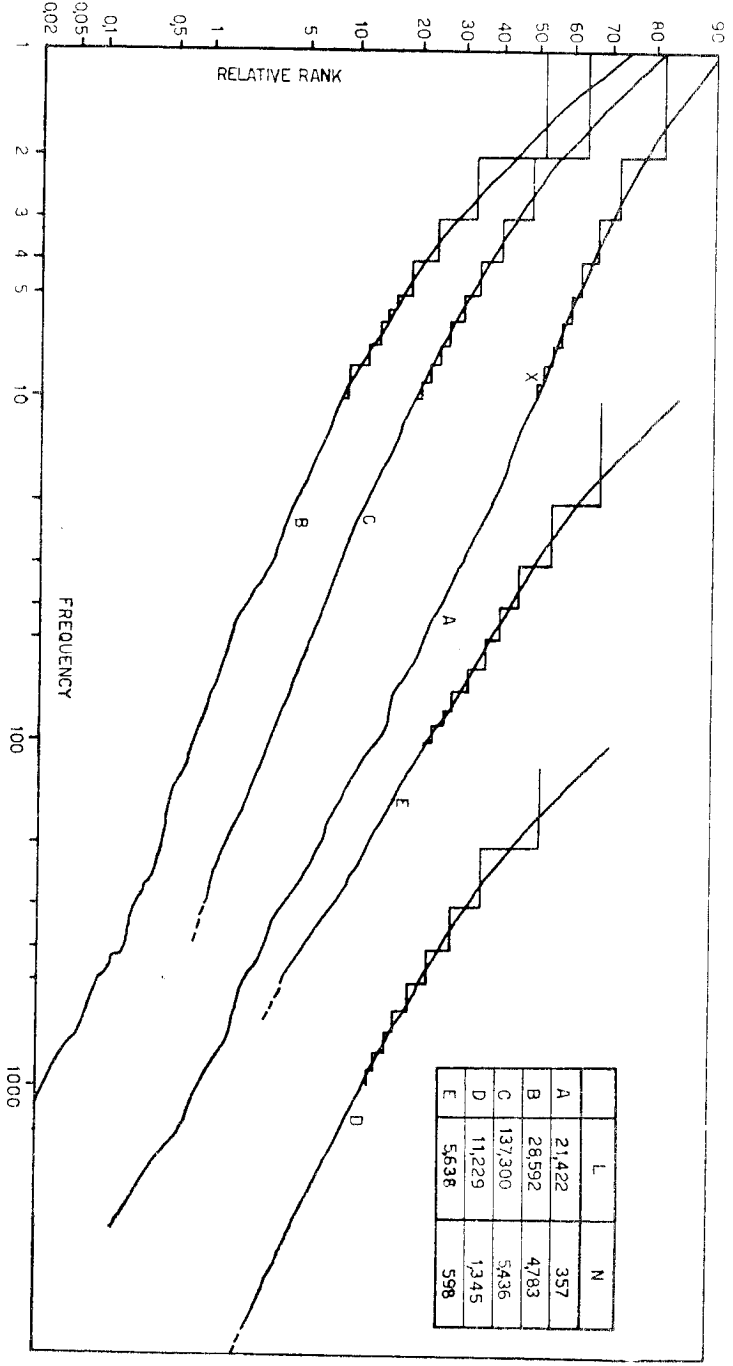


Fig. 3

becomes a straight line on logarithmic probability paper; this is checked for a few examples on Fig. 3, and it will be established herebelow that the deviations from linearity at low frequencies are quantitatively explained by the truncation effect.

The abscissae in fig. 3 are the absolute frequencies, as published in various sources ([6]), but curve E has been shifted by a factor 10 and curve D by a factor 100, to avoid overlapping. The table included in fig. 3 gives the extensions of the vocabulary (N) and of the text (L) mentioned in the source material. The discrete steps at the low frequency ends of the curves arise because the numbers of occurrences of the rarest words are necessarily small integers. Strictly, the usual definition of the rank-frequency relation would yield a continuous curve passing through the top points of the ladder, but the similar relation based on the complementary rank gives a curve passing through the bottom points. A unique smoothed continuous distribution curve can therefore only be defined by joining the vertical mid-points of the ladder, and this has been done in fig. 3. In particular, the first point of the smoothed distribution, corresponding to frequency 1, is $1 - N_1/2N$, where N_1 is the number of hapaxlegomena and N the total number of different words in the sample.

For curve A, the truncation effect is practically negligible because the statistics extends well above the mean. The slope of the straight line (abscissa interval corresponding to a decrease of the ordinate from 50 to 16%) defines the standard deviation as $\sigma = 2{,}8$ bits. From this value, and the value of N mentioned in the table, one can compute m by (13), and the corresponding median frequency is $L\, e^{-m}$; the value thus obtained is shown by a cross on fig. 3. Fig. 4 shows the curve of fig. 3A in bilogarithmic coordinates, and the dotted straight lines are Zipf's approximations at $x = x_a$ and $x = h$, the values of the slope being computed by (44), and the value of h by (15).

The truncation effect at the low frequency end will be discussed

([6]) Curve A is based on data for English function words (G.A. MILLER, E.B. NEWMAN, E.A. FRIEDMAN, « Length frequency statistics for written English », *Inform. and Control*, vol. 1 n° 4 pp. 370-389; Dec. 1958). Curve B corresponds to Russian words in the Captain's Daughter by Pushkin (H.H. JOSSELSON, *The Russian word count*, Detroit, 1953). Curves C (New testament) and D (St Mark) are based on R. MORGENTHALER, *Statistik des neutestamentlichen Wortschatzes*. Zurich 1958. Curve E is for French adjectives from all 8 tragedies of Racine (ref. note 4 p. 31).

on the example of curve B of fig. 3. The corresponding bilogarithmic representation of fig. 5 shows that the value of the initial slope

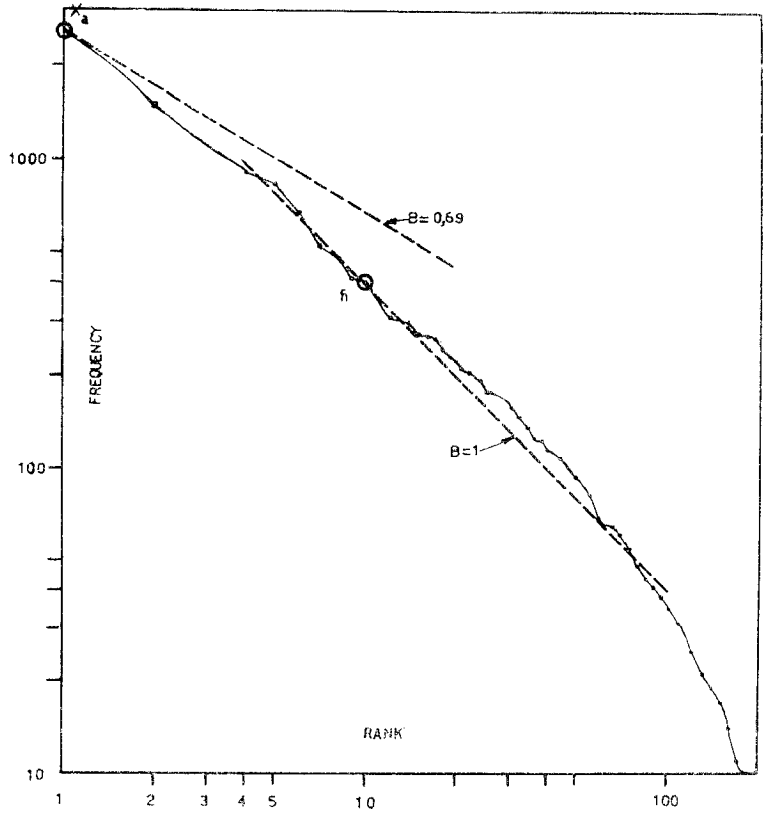


Fig. 4

is 0,48. Since the experimental value of x_a is 3,23 binits, (41) requires

$$0{,}48\ (m - 3{,}23) = \sigma^2$$

A second relation between m and σ is (30), for N is known and the experimental value of x is 10,9 binits. The solution of these equations is $m = 11{,}12$ binits $= 16{,}1$ bits and $\sigma = 2{,}8$ bits. The correction factor in (25) is

$$1/a = \Phi\left(\frac{x_b - m}{\sigma}\right) = 0{,}455$$

and this transforms curve B into curve B′ as shown on the right hand side of fig. 6. The cross corresponds to the computed value

for the curvature at the low frequency end.

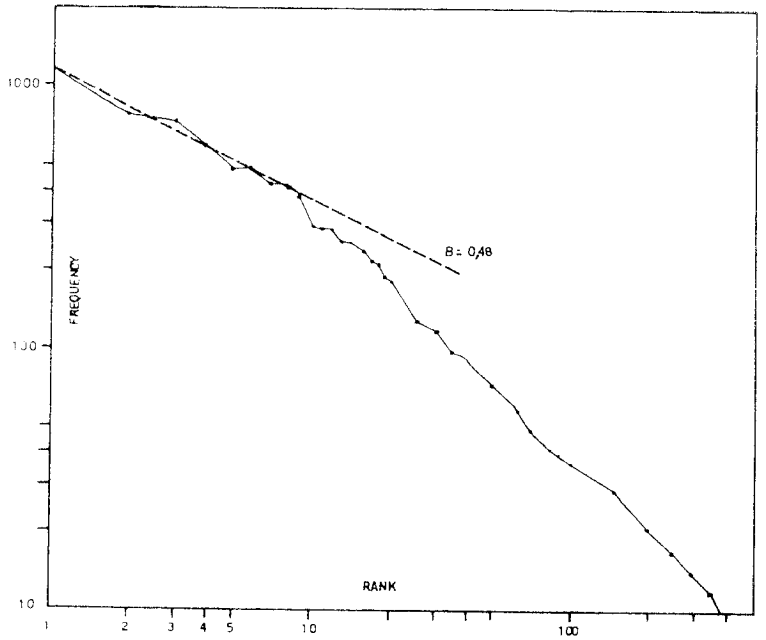


Fig. 5

The various examples of fig. 3 are represented in the left hand side of fig. 6 with an horizontal shift identical to the one discussed in passing from fig. 1 to fig. 2. The smoothed curves are still called A, B ... E, but the discrete steps have been omitted. It is apparent that all curves will become practically identical after correction for the truncation effect. This correction does not alter the slope of the linear part, and it is already obvious in fig. 3 that all distributions have practically the same slope. The common value seems to be 2,8 bits, which is the double of the value found for phonemes. The straight line in fig. 6 is the theoretical characteristic corresponding to $\sigma = 2{,}8$ bits.

If one accepts the normal distribution as the general law for words, the fact that Mandelbrot's or Zipf's laws are often satisfactorily confirmed would simply result from the enormous extension of the vocabularies combined with the limitation of many statistics well below the mean rank : for m large, B, as given (41), is constant

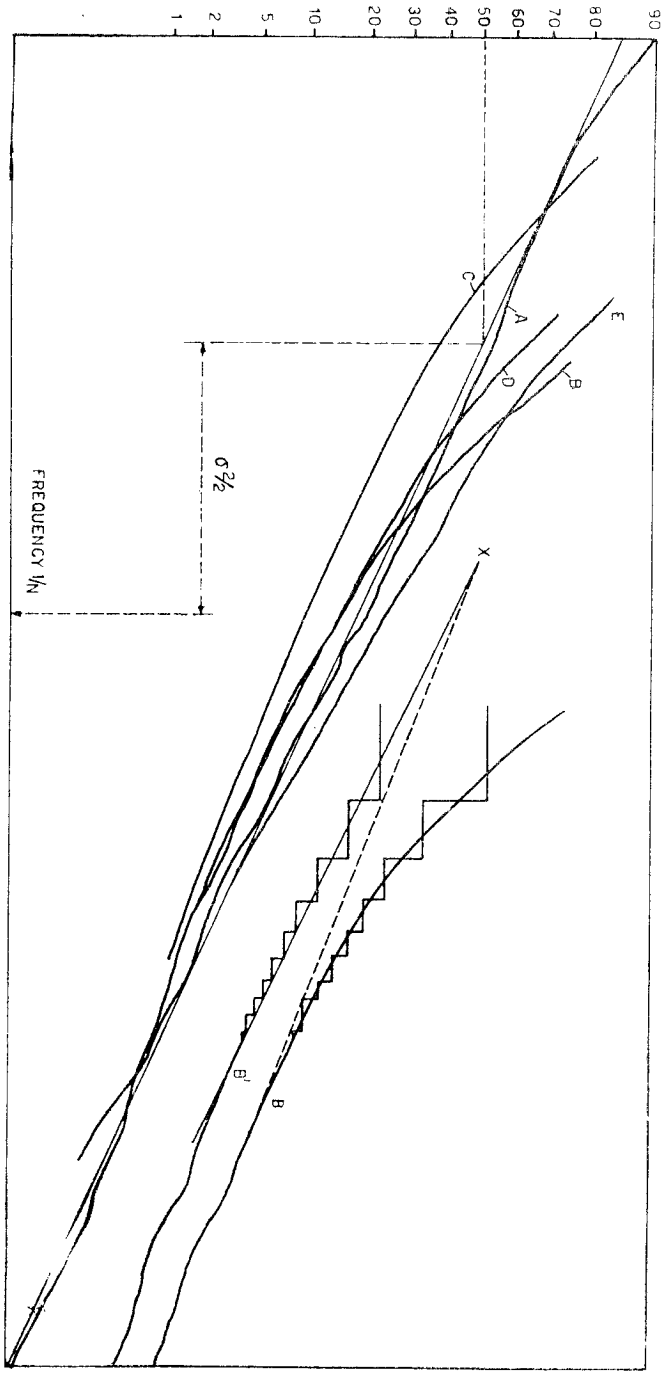


Fig. 6

would also explain the insistence of several authors in counting all inflected forms as distinct.

A number of theoretical models have been proposed by Mandelbrot to account for his law (7). The model based on the weakest hypothesis assumes that texts are separated into words by a randomly distributed ,,space'' symbol. No special theory is needed to explain a normal distribution, but it must be remarked that randomness has been shifted from the text to the vocabulary. For the high frequency tail of the distribution, where the saturation effect due to the finite extent of the vocabulary is still negligible, both explanations are equivalent, since the same stochastic model can be interpreted as yielding the text by a dilution of the dictionary, or the dictionary by a concentration of the text. But, for finite vocabularies a difference arises, because texts remain potentially infinite by hypothesis.

Finally, the constancy of the standard deviation for phoneme distributions on one hand, and for word distributions on the other, would suggest some common discrete substructure for both linguistic levels, but with a double number of degrees of freedom in the latter case.

Acknowledgement

The author is grateful to W. Croes and P.G. Neumann who criticized the manuscript.

(7) The publication of B. Mandelbrot, « Linguistique statistique macroscopique » in *Logique, langage et théorie de l information*, Presses Univ. Fr., Paris, 1957, gives a non-mathematical account of his theory, and a bibliography of the subject. See also the more recent contribution of B. Mandelbrot, in *Info and Control* vol. 2 pp. 90-99; april 1959.

L'énergie potentielle complémentaire dans les problèmes dynamiques. - Un principe de variation des accélérations

par

B. Fraeijs de Veubeke
Professeur aux Universités de Liège et de Louvain

Sommaire

Tout comme le principe de variation des déplacements en élasticité, le principe de Hamilton peut être transformé en un principe canonique autorisant des approximations indépendantes sur les déplacements et sur les forces de liaison.

On en déduit un principe dual à celui de Hamilton lorsque les équations d'équilibre dynamique sont exactement satisfaites. Pour des liaisons indépendantes du temps et une énergie cinétique fonction des seules vitesses ce principe dual prend la forme plus élégante d'un principe de variation des accélérations.

Ces idées sont d'abord exposées pour les systèmes à nombre fini de coordonnées lagrangiennes et appliquées, à titre d'exemple, à la recherche de valeurs approchées à la période d'oscillation d'un pendule. Elles sont ensuite étendues au cas de l'élastodynamique linéaire et appliquées à un cas simple de vibrations d'une poutre hyperstatique.

1. Généralisation du principe de Hamilton

Soient $q_r (r = 1, 2 \ldots n)$ des coordonnées généralisées permettant de décrire la configuration d'un système holonome à n degrés de liberté. Le principe de Hamilton (1)

$$\delta \int_{t_1}^{t_2} \left[T(q_r, \dot{q}_r, t) - V(q_r) \right] dt = 0 \qquad (1)$$

$$\delta q_r = 0 \text{ pour } t = t_1 \text{ et } t = t_2 \qquad (2)$$