Glottometrics 5, 2002

To Honor G.K. Zipf

RAM - Verlag

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –5 (2002) –. – Lüdenscheid: RAM-Verl., 2002
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 5 (2002)
ISSN 1617-8351

Contents

Zipf´s law and its modification possibilities

Victor Kromer[1]

Abstract. In this paper we consider the possibilities of known Zipf-Mandelbrot canonical law modifications. The proposed modifications explain the behavior of the right tail of the distribution and the presence of a deflection in the central part of the distribution (a crater). It is shown that the average word information load is invariant to the sample heterogeneity and that the proposed usage measure "places" the words more correctly with regard to their "importance".

Keywords: crater, packing density, resolving power, usage measure, Zipf´s law

At present many empirical and theoretical expressions for describing the relationship between the word frequency (absolute or relative) or the word probability and word number in a sequence ranked on the decrease of frequency are known. The most simple relationship of a similar kind is called by right Zipf´s Law (classical one-parameter Zipf's distribution) and relates word rank *r* and word probability p_r:

$$p_r = \frac{K}{r}, \tag{1}$$

where *K* is the coefficient of proportionality. This relationship, theoretically justified by B. Mandelbrot as a corollary of an optimal coding process, is titled as Zipf-Mandelbrot canonical law and has a more generalized nature:

$$p_r = \frac{K}{(r+B)^{\gamma}}, \tag{2}$$

where *B* and γ are distribution parameters. Later on Ju.K. Orlov showed that the relationship of type (2) at $\gamma = 1$ (Orlov, 1978, pp. 85, 89) holds for the so-called "optimum" or "Zipf's" samples, basically representing complete texts of separate literary productions. The size of an optimal sample, as a rule, lies within a rather narrow range giving no way to use formula (2) for describing mixed or truncated samples. The researchers offered various corrections to the formula (2), in an effort to describe similar samples, strongly differing in their frequency structure from the optimum sample structure. Any correction increases the number of parameters in the resulting formula and therefore allows achieving a better agreement between the empirical relationship and the theoretical one. However, the degree of agreement cannot be a decisive factor in favor of one or another theoretical distribution, since: (i) Formulas like (2) and more sophisticated ones describe in fact word probability distribution not in a sample, but in the general population. The actual word frequency distribution of the sample is determined by Poisson's law and can be described by the formula which is appropriate for describing word frequencies distribution in the general population, only by an

[1] Address correspondence to: V. Kromer, Viljujskaja ul., 28, NGPU, Novosibirsk, 630126, Russia. E-mail: applied@nspu.ru. URL: http://kromer.newmail.ru.

artificial modification of formula parameters in order to account for the peculiarities of empirical distribution caused by Poisson's law action. (ii) With an increase of the sample size any theoretical distribution, even the best one, should be rejected by goodness-of-fit test. The cause of discrepancies consists in the fact that in large samples many subsidiary conditions come along which are not taken into account by the theoretical distribution. In samples of a smaller size manifestation of the same neglected factors could be explained by fluctuations.

In view of the aforesaid, the kinds of corrections, which take into account actual cognitive processes, which are related to text generation and text selection and have linguistic meaning, are the most valuable ones (considering their pithiness and prognostic strength). The "right tail" of the distribution, as it is called – the range of high ranks (low frequencies) is of special interest for researchers. It is the practice to describe this area by the spectral analog of rank distribution, called Pareto's law:

$$m_F = \frac{A}{F^{\alpha}}, \tag{3}$$

where m_F is the number of sample words, having frequency *F*, and *A* and α are distribution parameters.

The relation between distributions like (2) and (3) was repeatedly mentioned in the literature (Tuldava, 1987, p. 86). The expressions (2) and (3) are two forms of the same relationship, with parameter values related by the expression $\alpha = 1 + \frac{1}{\gamma}$ (and accordingly $\gamma = \frac{1}{\alpha - 1}$); (Tuldava, 1987, p. 88). The value $\alpha = 2$ of Pareto's law parameter corresponds to the value $\gamma = 1$, offered by Zipf. The given values are the most typical ones for linguistic samples. The value of parameter γ, falling in the range of 0.8–1.2 (depending on the language, genre, author etc.) corresponds to the α-parameter varying within the range of 1.83–2.25. The value $\alpha = 1$ corresponds to the inversely proportional relationship between frequency and the number of words with the given frequency. Considering the number of words with the given frequency to be a frequency (frequency of frequencies) and ranking groups of words with identical frequency, we can assign rank 1 to the most numerous group ($F_1 = 1$), rank 2 to the next group ($F_2 = 2$) etc.; then expression (3) appears to be identical to expression (1) providing *F* is the rank, and the α-value is equal to 1. (This condition is met only in the range of not very high frequencies, when all frequency values are presented in the sample).

As at $\alpha \to 1$ the corresponding value of $\gamma \to \infty$, the rank-frequency relationship cannot be described by power law relation (2) and another kind of expression is required. It can be obtained by integrating expression (3). Let us assign the rank $r^{(F)}$ to the first word in rank distribution with given frequency *F*, and rank $r^{(F-1)}$ to the first word with frequency (*F*-1). (From here on *r* is the traditional word rank, i.e. word number in the list of words ranked on decreasing frequency). The ratio between the increment of rank and the increment of frequency is $\frac{\Delta r}{\Delta F} = \frac{r^{(F-1)} - r^{(F)}}{(F-1) - F} = \frac{m_F}{-1} = -m_F$. Considering *r* and *F* as continuous variables and changing from differences to differentials, we obtain $\frac{dr}{dF} - m_F$. At $\alpha = 1$ we obtain

$$\frac{dr}{dF} = -\frac{A}{F}, \tag{4}$$

whence it follows that

$$r = -A \ln F + C, \tag{5}$$

where *C* is the constant of integration. Solving equation (5) for *F*, we obtain

$$F = \frac{e^{C/A}}{e^{r/A}}. \tag{6}$$

Passing to probabilities, we obtain

$$p_r = \frac{M}{e^{r/A}} = \frac{M}{d^r}, \tag{7}$$

where $d = e^{1/A}$ is a constant, slightly exceeding 1, and *M* is the normalizing coefficient.

Expression (7) is frequently used for describing rank-frequency distribution of small inventory units (alphabet signs, phonemes, DNA codons etc.). Expression (7) is unsuitable for describing lexical frequency structure of texts. At the same time expression (7) has an interesting peculiarity. Let's find the information load (negative entropy) of a word (measured in nits[2]) at rank *r* providing the word distribution is in accordance with (7):

$$I_r = \ln\frac{1}{p_r} = \ln\frac{d^r}{M} = r \ln d - \ln M. \tag{8}$$

The first derivative of I_r with respect to *r* makes $\frac{dI_r}{dr} = \ln d$. This value has the meaning of information distance between adjacent units of the rank distribution. The inverse is the formerly defined (Kromer, 1997a, p. 29) packing density $D = \frac{1}{dI_r/dr} = \frac{1}{\ln d}$ of units (words). As is obvious, the vocabulary packing density for distribution (7) is constant. Zipf's law in its canonical form (2) does not limit this characteristic feature with increasing word rank, as $D_Z = \frac{r+B}{\gamma}$, which can be proved by finding logarithm and differentiating expression (2).

There is a supposition that psychophysical relations, valid for elementary sensual acts (sight, hearing, sense of touch) can be extended to more intricate aspects of mental activity, for example to processes of unfamiliar vocabulary perception and preservation (Kondrat'eva, 1972, p. 40). Let's extend these relations to processes of text generation and text perception (as in the course of text generation the author anticipates the possibility of text perception by the recipient), as well as to processes of texts selection (compiling of reading-books, anthologies, text corpora etc.) (Kromer, 1999a). Now suppose there exists a hypothetical expression, based on (2) and expressing rank-frequency relation in view of some limitations. It is well known that quantitative properties of human memory (and, maybe, collective memory properties) are limited. In particular, maximum word information load is limited (Piotrovskij et al., 1997, p. 94).

The suggestion has been made that a person's ability to resolve individual word probabilities on the allocated information space is also limited, and this feature is characterized, as we suppose, by the maximum vocabulary packing density. This performance

[2] 1 nit = $\log_2 e$=1.443 bit

has the meaning of a resolving power (RP), similar to RP of sight, hearing etc. The limitless RP of Zipf′s law encounters restrictions imposed by a particular language, sublanguage, idiolect. A crude expression relates RP of a complex system and its separate components (Prochorov, 1984, p. 615):

$$\frac{1}{R_s}=\sum_{i=1}^{n}\frac{1}{R_i}, \tag{9}$$

where n is the number of system components, R_s is system RP, and R_i is RP of the i^{th} system component. The above expression at $n=2$ is valid, in particular, for RP of an optical device made up of an optical system (lens) and a receiver (for example, a light-sensitive layer). Let's designate as D_L the limiting vocabulary packing density of a particular language. Then the following expression would be correct:

$$\frac{1}{D_s}=\frac{1}{D_Z}+\frac{1}{D_L}, \tag{10}$$

where D_Z stands for vocabulary packing density, given by Zipf′s law in its canonical form (2). D_s is the vocabulary-packing density of the complex system "canonical Zipf′s law – particular language". Let's pass to corresponding information distances:

$$\frac{dI_s}{dr}=\frac{\gamma}{r+B}+\frac{1}{D_L}. \tag{11}$$

By integrating (11) we can find I_s – word information load in accordance with the sought-for hypothetical expression, reflecting rank distribution of the complex system:

$$I_s=\int\frac{dI_s}{dr}dr=\int\left(\frac{\gamma}{r+B}+\frac{1}{D_L}\right)dr=\gamma\ln(r+B)+\frac{r}{D_L}+C, \tag{12}$$

where C is the integration constant. As $I_s=-\ln p_s$, where p_s stands for word probability in the complex system, we can deduce the expression for p_s by transforming:

$$p_s=e^{-I_s}=e^{-C}\,(r+B)^{-\gamma}\left(e^{-1/D_L}\right)^r. \tag{13}$$

Substituting constants e^{-C} for K and e^{-1/D_L} for z, and designating probability as p_r, we obtain the following expression:

$$p_r=\frac{Kz^r}{(r+B)^\gamma}. \tag{14}$$

This expression is known as canonical law with J. Woronczak's correction (Woronczak, 1967, p. 2226). The additional parameter $z<1$ assures faster decreasing of function (14) in comparison with function (2) and specifies behavior of the distribution "right tail". Thus, Woronczak's correction gains confirmation in the context of the proposed supposition about extending of known psychophysical relations to intricate aspects of mental activity.

Besides "right tail", the empirical distribution is characterized by one more peculiarity, labeled as "crater" (Kromer, 1997a, p. 22). The mentioned peculiarity is most common for mixed samples. The availability of the crater in the distribution can be revealed on the graph of rank–frequency distribution, plotted on double logarithmic axes. However, the crater availability is more explicit by using a special form of data displaying, if the logarithm of the rank is plotted along the axis of abcissas and the product of the rank by relative frequency along the axis of ordinates. This form of data displaying (Nešitoj, 1987) has some advantages, such as graph compactness in vertical direction (the second bisectrix of the traditional coordinate system "logarithm of rank – logarithm of frequency" is mapped in the new coordinate system as a horizontal line). As an example we calculated for the text corpus of Frequency Dictionary of Russian Language – FDRL (Zasorina, 1977), comprising 1,056,382 running words, the value of $G = r \cdot f_r$, where *r* stands for word rank, and f_r stands for relative word frequency, computed as ratio of absolute frequency to the sample size. The calculation results are depicted in graphic form in Figure 1.

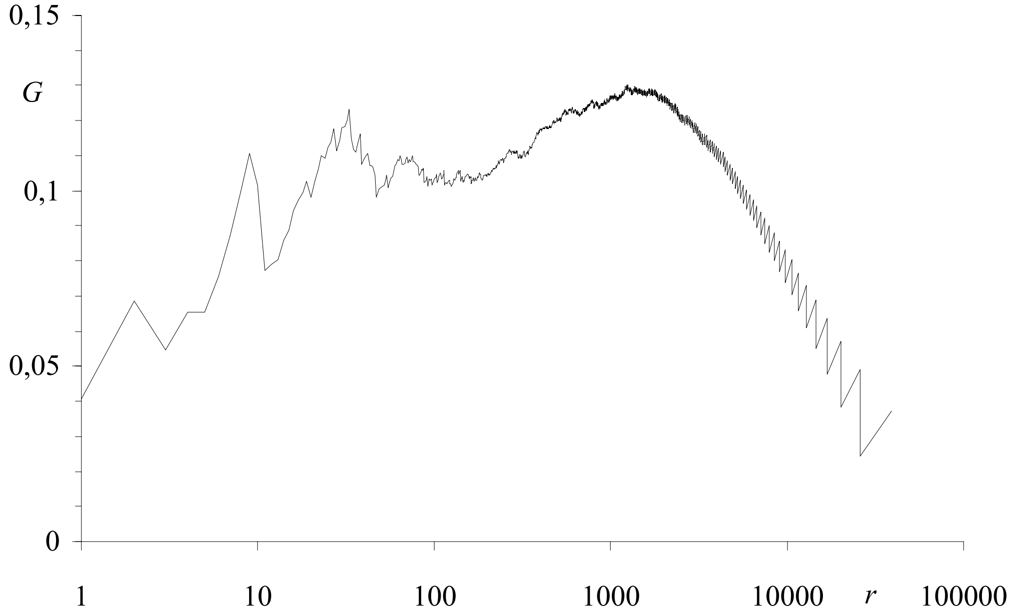


Fig. 1. $G = rf_r$-value as a function of rank *r* for FDRL

From the graphic data it can be seen that the graph of $G = rf_r$ for FDRL is characterized by the presence of a crater (deflection in the central part of the distribution). The similar deflection is also characteristic for other mixed samples (text corpora) distributions. According to Nešitoj (1987, p. 125) unimodalness of such a curve is an evidence of sample homogeneity. The research of horizontal word–frequency distribution on genre groups of texts has been carried out in effort to reveal the mechanism for crater formation in the distribution. The authors of FDRL distinguished 4 genres (styles): newspaper writing, drama texts, scientific and publicistic texts, belles lettres (fiction). The coefficient of skewness of the horizontal distribution was calculated for the 11,576 most frequent words of FDRL (with frequency equal and exceeding 7). The smoothing of the coefficient of skewness was carried out with the intent to get generalized characteristics of particular frequency spectrum parts. The frequency distribution proves to be right-handed and can be characterized by a skewness coefficient of about 0.8. The sample frequencies were replaced by a special function of frequency (SFF) – the sum of first F_j terms of harmonic series $1+\frac{1}{2}+\frac{1}{3}+\ldots+F_j$, where F_j is the frequency of a particular word (lemma) in the sample of the genre under consideration (Kromer, 1997b). The value of SFF, equal to 0, corresponds to $F_j = 0$ by definition. The skewness of SFF distribution (after smoothing) appears to be very close to 0. Since the sum of harmonic series is expressed by the psi-function (the logarithmic derivative of an Euler's integral of the second kind) as $\psi(F_j+1)+C$, where $C = 0{,}5772..$ is Euler's constant, we can say that the horizontal distribution of psi-function of frequency increased by unity is symmetric on the average.

Examination of SFF excess coefficient reveals that the horizontal SFF distribution is approximately uniform at the beginning of rank distribution (up to ranks about 600). Then the distribution peakedness increases, i.e. the density of SFF-values increases at the center of horizontal distribution, and as a first approximation it is possible to consider the SFF distribution as a normal (Gaussian) one. This conclusion, regarding special features of word probabilities in horizontal distribution, goes back to Arapov's (1988, p. 54) suggestion to consider language as a set of coordinated vocabularies. The sum of harmonic series over the range of sufficiently high frequencies can be given by approximate expression

$$\sum_{k=1}^{F_j} \frac{1}{k} \approx \ln F_j + C. \tag{15}$$

Subtracting from the right and left sides of equation (15) the constant $\ln N$ (where *N* stands for genre sample size) and rearranging Euler's constant *C* to the left equation side, we obtain:

$$\sum_{k=1}^{F_j} \frac{1}{k} - (\ln N + C) = \ln F_j - \ln N = \ln \frac{F_j}{N} = \ln f_j, \tag{16}$$

where $f_j = \frac{F_j}{N}$ is the relative frequency of the word with absolute frequency F_j. The distribution of the left side of equation (16), differing from (15) by the constant value $(\ln N + C)$, would be normal, providing the distribution (15) is normal. Based upon the normality of horizontal SFF distribution, with high frequencies the logarithm of the relative frequency is distributed normally as well, i.e. the relative frequency of words has logarithmic-normal (lognormal) horizontal distribution. What was said regarding horizontal distribution evidently can be extended to word probabilities. The essence of the model offered in (Kromer, 1997c, p. 20) is the assumption that it is not the word probability which is distributed in accordance with Zipf-Mandelbrot law, but the exponential function of the mathematical expectation of word probability logarithm:

$$\exp(M(\ln p_r)) = \frac{K}{(r+B)^{\gamma}}, \tag{17}$$

where $M(\ln p_r)$ is the mathematical expectation of word probability logarithm. As for the probability logarithms themselves, their horizontal distribution is characterized by dispersion σ_{hor}^2, where σ_{hor} is the standard deviation of probability logarithm, and it increases with increasing word rank. The offered model is constructed as a result of research on a text corpus, comprised of 4 groups of various genre texts. In reality the number of genre groups could be enhanced without bound, as the words probabilities in horizontal distribution do not assume discrete values, but are smeared-out in some range, and σ_{hor} accounts for probability scatter of the word under consideration.

From the above reasoning, the mathematical expectation of word probability is determined by this expression (Prochorov, 1982):

$$M(p_r) = \exp\left(M(\ln p_r) + \frac{\sigma_{hor}^2}{2}\right) = \frac{K}{(r+B)^{\gamma}} \exp\left(\frac{\sigma_{hor}^2}{2}\right). \tag{18}$$

It is accepted in the model that σ_{hor} takes its minimum value at the beginning of the rank distribution, then these values increase with increasing word rank, establishing on some rank their limiting value, and then keep constant up to the end of rank distribution. The term $\exp\left(\frac{\sigma_{hor}^2}{2}\right)$ of expression (18) accounts for the approximate effect of mixed sample heterogeneity having regard to the offered assumption. In the general case

$$p_r = \frac{K}{(r+B)^\gamma} f(r), \qquad (18')$$

where $f(r)$ is a nondescending function of r, varying from 1 to s, where $s>1$ is the maximum value of the function. (Despite the fact that the minimum value of σ_{hor} is not equal to 0, the minimum value of $f(r)$ can be equal to 1 at the expense of the correction of the normalizing coefficient K).

The chief drawback of the representation of the distribution form offered by Nešitoj (1987) is the curvature at $\gamma \neq 1$ in the graphs of dependencies, mapped by straight lines by the traditional representation. The said drawback can be eliminated by plotting along the ordinate axis the logarithm of rank multiplied by the frequency or the sum of rank and frequency logarithms. Thus, the offered representation form is the traditional one, such that the ordinates of the dependence points are increased by the rank logarithm (Kromer, 1999b, p. 16). It is also possible to plot as ordinates the $S_r^{(t)} = \ln\left(F_r(r+B)^\gamma\right)$ value, where F_r is the word frequency, r is the word rank, and B and γ are the parameters of Zipf-Mandelbrot distribution for the initial part of the distribution. In this case the initial part of the distribution is plotted as a horizontal straight line. A growth of the dependence is observed in the crater domain, reflected by member $f(r)$of equation (18'). Further a horizontal plateau or decreasing part reflecting the increase of γ on high ranks and described by Woronczak's correction can follow. In practice the initial part of the distribution is polygonal at the expense of fluctuations, and the distribution graph is not plotted in compliance with previously known parameters γ and B, but those parameters are fitted with respect to the best straightening of the initial part of the distribution (for which parameter B is responsible) and its horizontal leveling (for which parameter γ is responsible). Such a graph for FDRL using estimated parameters $\gamma = 1.032$ and $B = 2.42$ is plotted in Figure 2 (thin line). The parameters are estimated for the initial part of the distribution (words ranked from 1 up to 100). We use the concept "initial part of the distribution" without its explicit determination. Later on in this paper we will return to this concept and set a procedure providing estimation of the terminal rank of the initial part.

The question arises of whether there exists a word feature invariant to the sample heterogeneity. By virtue of the assumption that the exponential function of mathematical expectation of word probability logarithm is distributed in line with the canonical law (17), the exponential function of increased word average information load (as word information load is equal to the inversed logarithm of probability) is distributed in accordance with the same law. As a consequence,

$$\exp(-\bar{I}_{\text{inf}}) = \frac{K}{(r+B)^\gamma}, \qquad (19)$$

where $\bar{I}_{\text{inf}}$ is the average word information load. It is worthy of note that $\bar{I}_{\text{inf}}$ can be estimated from the word rank using equation (19), where parameters K, B and γ are estimated for the initial part of rank-frequency distribution. The total word frequency in a mixed sample (and in a sense all samples can be considered as mixed ones) can be used only for rank estimating, but not for direct estimating of average information load. The word information load, computed on the formula $I_{\mathrm{inf}} = -\ln p_r$ (where p_r is the probability of the word, ranked r) is

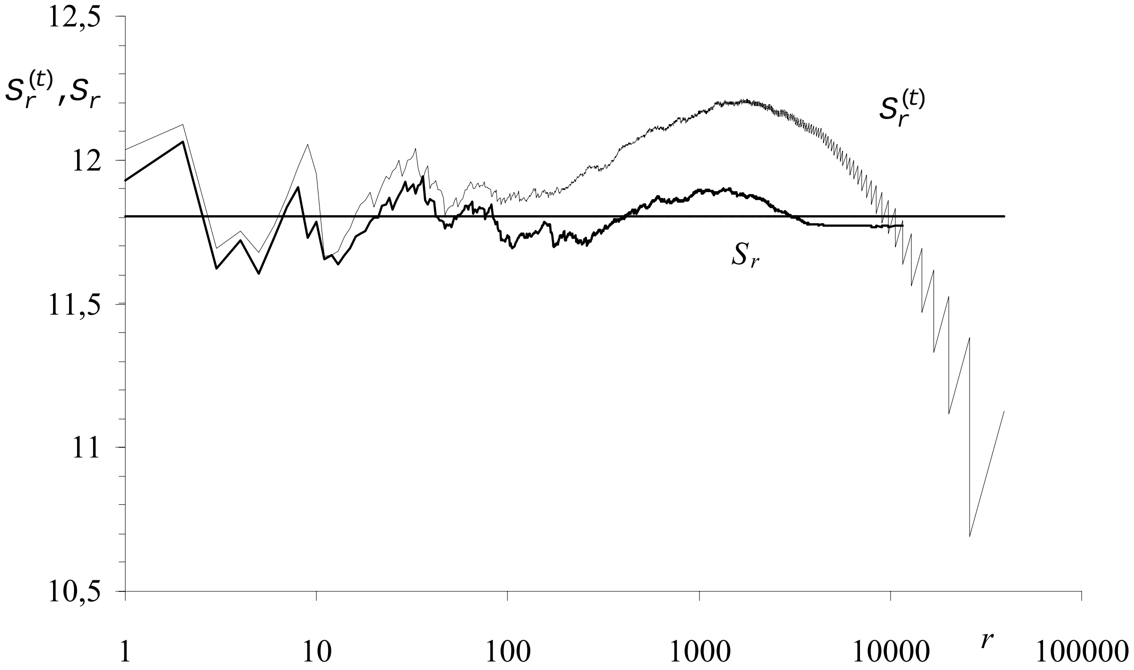


Fig. 2. $S_r^{(t)}$ and S_r dependencies as functions of rank r for FDRL

generally underestimated, i.e. $I_{\mathrm{inf}} \leq \bar{I}_{\mathrm{inf}}$. When evaluating $\bar{I}_{\mathrm{inf}}$, representing the language as a whole on empirical data, difficulties emerge, as the mixed sample (the text corpus, representative for the language as a whole), should be divided into constituent homogeneous texts, and the information load should be evaluated for each corpus word for every constituent text. As not all corpus words are present in each particular text, it is not possible to evaluate the information load of text zero-frequency words. In addition, the estimated information load of words with frequency less than 5–10 is unreliable. The method of sparse data smoothing by way of probability redistribution (additive smoothing method) exists, that means giving some small probabilities to zero-frequency words and correcting the probability of low frequency words at the cost of decrease of the probability of high frequency words. This method consists in adding some constant k to all the frequencies (including the zero frequencies), and the probability $p^{(i)}$ of word occurrence in i^{th} text is estimated relying on the maximal likelihood estimation method:

$$p^{(i)} = \frac{F^{(i)} + k}{N_i + kL}, \tag{20}$$

where $F^{(i)}$ is word frequency in i^{th} text, N_i is size of the i^{th} text in running words, L is total number of distinct words in the text corpus (or, what we consider to be the same, number of potential distinct text words including zero-frequency words). For $k=1$ the method is known as Laplace's law, for $k=0.5$ – Lidstone's law (Nivre, 2000, p. 4). Dwelling on the last alternative, the word information load can be estimated from the corrected word probability $p^{(i)} = \frac{F^{(i)} + 0.5}{N_i + 0.5L}$:

$$I_i = \ln \frac{1}{p^{(i)}} = \ln \frac{N_i + 0.5L}{F^{(i)} + 0.5} = \ln(N_i + 0.5L) - \ln(F^{(i)} + 0.5). \qquad (21)$$

The average word information load in the corpus consisting of *n* texts yields

$$\overline{I}_{\text{inf}} = \frac{\sum_{i=1}^{n} I_i}{n} = \ln(N + 0.5L) - \frac{\sum_{i=1}^{n} \ln(F^{(i)} + 0.5)}{n}. \qquad (22)$$

Expression (22) is deduced on condition that all the corpus texts are equally sized as *N* running words.

An approximation to the sum of harmonic series closer than in expression (15), based on the limit $\lim_{x\to\infty}\left[\exp\left(\sum_{k=1}^{x}\frac{1}{k} - C\right) - x\right] = 0{,}5$ is known:

$$\sum_{k=1}^{F^{(i)}} \frac{1}{k} \approx \ln(F^{(i)} + 0.5) + C. \qquad (23)$$

Let's find the sum of SFF for each corpus word, using expression (23):

$$\sum_{i=1}^{n} SFF = \sum_{i=1}^{n} \ln(F^{(i)} + 0.5) + nC. \qquad (24)$$

As is obvious by comparison of expressions (24) and (22), the average word information load (providing sparse data is smoothed in accordance with Lidstone's law) and the SFF sums for words are related by linear transformation, so if required, the SFF sum offers an alternative to the average information load, and that removes the problem of text zero-frequency words, as SFF is equal to 0 for them.

Let's test the assumption that the average information load of FDRL corpus words, represented by the SFF sum, is invariant with respect to the sample heterogeneity. The value

$$S_r = \frac{\sum_{j=1}^{4}\left[\psi(F_j + 1) + C\right]}{4} + \gamma \ln(r + B) + q, \qquad (25)$$

which is the SFF sum averaged over 4 genre samples, increased by $\gamma \ln(r + B)$ for the purpose of dependence straightening and horizontal leveling, and by some constant *q* to normalize 2 dependencies in Figure 2 (which means in this context the maximum alignment of initial parts of dependencies for the purpose of comparison). Before computing the SFF sum the size of each genre samples was corrected with respect to size inequality of genre samples by multiplying sample sizes by the normalizing coefficient (1.050 for newspaper texts, 0.919 for drama text, 1.059 for scientific and publicistic texts and 0.984 for fiction).

Parameters γ and *B* of expression (25) are also estimated in order to provide the best straightening and horizontal leveling of the initial part of dependence S_r as a function of *r*. In general the values of those parameters differ from the like parameters of the canonical law, as the distributions of $\ln F_r$ and SFF sum are different ones, and each of them is characterized by its own set of parameters. Let's show that distribution parameterization provides a way of estimating the size of the initial part of the distribution. The words ranked from 25 up to 500 with a step of 25 may be considered as the potential right boundary *R* of the initial part of the distribution. We can use the method of least square deviations for estimating the parameters of the trend line of the dependencies $S_r^{(t)} = \ln F_r (r+B)^\gamma$ and S_r (formula 25). The trend line is regarded as a parabola, described by the equation

$$y = ax^2 + bx + c, \qquad (26)$$

where $x = \ln r$ is the abscissa, and *a*, *b* and *c* are coefficients. Let's consider parameters γ and *B* of the dependencies $S_r^{(t)}$ and S_r as unknown ones, to be found proceeding from conditions $a = 0$ (resulting in the linearity of the trend line) and $b = 0$ (resulting in the horizontality of the trend line). The estimated values of γ and *B* for 2 dependencies under consideration are presented in Table 1.

Table 1

Estimated values of γ, *B* and Δγ for dependencies $S_r^{(t)}$ and S_r

R		25	50	75	100	125	150	175	200	225	250
$S_r^{(t)}$	γ	0.805	0.916	0.988	1.032	1.051	1.053	1.052	1.047	1.039	1.026
	B	0.45	1.29	1.92	2.42	2.66	2.69	2.68	2.60	2.45	2.22
S_r	γ	0.829	0.901	0.973	1.045	1.093	1.097	1.091	1.097	1.095	1.098
	B	0.41	0.94	1.53	2.31	2.92	2.97	2.89	2.98	2.94	3.01
Δγ		0.024	-0.015	-0.015	0.013	0.042	0.044	0.039	0.050	0.056	0.072
R		275	300	325	350	375	400	425	450	475	500
$S_r^{(t)}$	γ	1.012	1.003	0.998	0.993	0.986	0.977	0.968	0.962	0.957	0.952
	B	1.97	1.79	1.69	1.60	1.45	1.26	1.10	0.98	0.87	0.76
S_r	γ	1.099	1.091	1.083	1.074	1.064	1.056	1.047	1.040	1.033	1.027
	B	3.02	2.86	2.69	2.50	2.29	2.10	1.93	1.75	1.60	1.46
Δγ		0.087	0.088	0.085	0.081	0.078	0.079	0.079	0.078	0.076	0.075

According to the data of Table 1, parameters γ and *B* vary in a regular way in dependence of *R*. Dependence $S_r^{(t)}$, associated with frequency, reveals a local maximum of γ at ranks about 125–175. According to Kromer (1997a, p. 31) the word ranked 136 is the center of the crater. The parameters γ and *B* of SFF sum distribution (formula (25)) reveal different values for the corresponding values of *R*, but they also vary in a regular way. The list of words, ranked by frequency, differs from the similar list, ranked on SFF sum, and values of γ in dependence of *R* differ for 2 distributions under consideration. The difference of γ-values for 2 distributions (Δγ) is of principal interest for us. It is seen from Table 1 that this difference increases monotonically, reaching saturation. Precisely this difference characterizes the behavior of function $f(r)$ (formula 18') reflecting disparity between the frequency and SFF sum at the expense of an increase of the variance of horizontal distribution frequency. According to the data of Table 1 the initial part of rank-frequency distribution includes about 100 words, as the difference between 2 distributions (Δγ) increases at higher ranks. The graph of dependence S_r (SFF sum, averaged over 4 genre samples) is plotted in Figure 2 (heavy line) according to parameters γ and *B* from Table 1 estimated at rank 100. Two graphs in Figure 1 are normalized, i.e. their initial parts are aligned as much as possible, which requires $q = 0.862$ (formula 25).

From comparison of 2 graphs it follows that the average SFF sum for the FDRL text corpus is not completely invariant to the sample heterogeneity (the differential between the initial level and the saturation level makes 0.25 logarithmic units (LU) for the 1st graph, and 0.09 LU for the 2nd one). The retained nonzero differential is attributable to the fact that the genre samples of FDRL are not completely homogeneous texts with Zipf distribution, but mixed samples with smaller heterogeneity degree, than that of the total text corpus. The graphs $S_r^{(t)}$ for all 4 genre samples of FDRL are plotted in Figure 3. Parameters γ and *B* for the same 4 dependencies are presented in Table 2. The initial parts of the dependencies were also considered as 100 words long. The dependencies were also normalized by leveling them to the identical level.

Table 2

Parameters γ and *B* for $S_r^{(t)}$ dependencies for 4 genre samples of FDRL

Genre	Newspaper texts	Drama texts	Scientific and publicistic texts	Fiction
γ	0.933	1.028	0.877	1.043
B	1.49	1.84	0.98	2.52

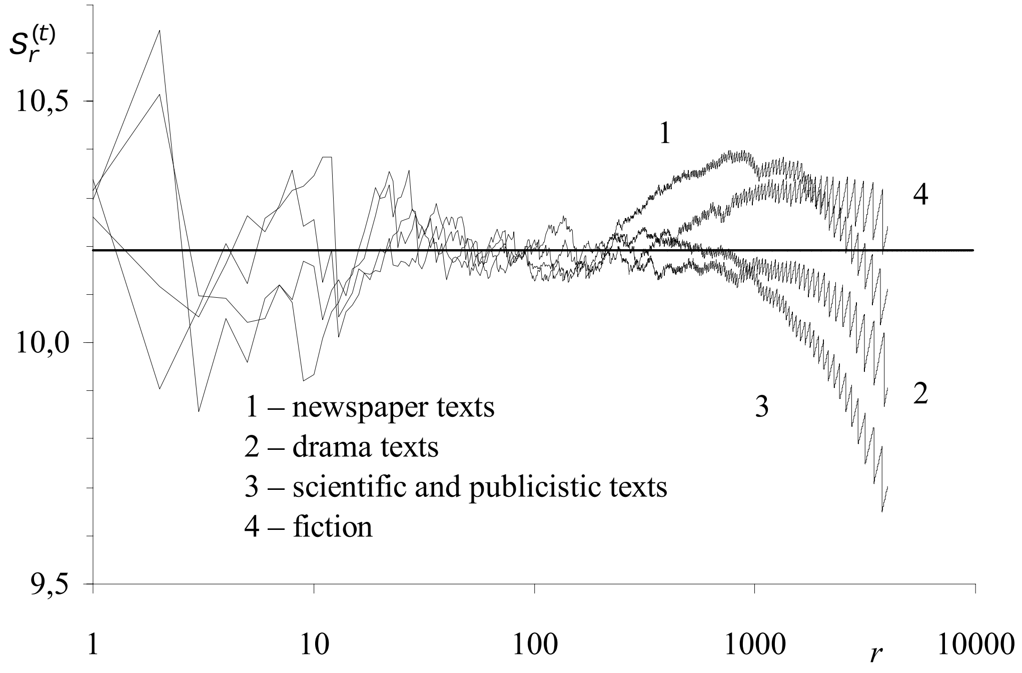


Fig. 3. Dependence $S_r^{(t)}$ as a function of *r* for 4 genre samples of FDRL

The degree of heterogeneity of FDRL genre samples, estimated by the level differential of the graphs, ranges from 0 up to 0.19 LU, i.e. the level differential of SFF sum graph for the text corpus could be completely attributed to the genre samples' inhomogeneity. According to Figure 3 the genre samples of drama texts and scientific and publicistic texts are the most homogeneous, the genre samples of newspaper texts and fiction are the most inhomogeneous. The parameters γ and *B* of all distributions were estimated using a strictly formalized procedure, which lends credence to the revealed empirical regularities and to the conclusions made on their basis. It is also well to bear in mind that graphs in Figures 2 and 3 were plotted with respect to the ranks of word frequency or word SFF sum, and generally different words are in correspondence with the particular rank. This brings up the question which of the two distributions under consideration (depending on frequency or SFF sum, i.e. on the average word information load) "places" the words more correctly with regard to their "importance".

A usage measure, based on extending the known psychophysical Weber-Fechner's law on the process of text perception, was offered in (Kromer, 1998). It was proposed to accept the expression

$$U_R = \sum_{j=1}^{n} \left(\psi(F_j + 1) + C \right) \tag{27}$$

as a usage measure, where n is the number of texts in the corpus, and F_j is the word frequency in the j^{th} text. From the aforesaid, it might be assumed that ranking the vocabulary by usage measure U_R is equivalent to ranking with respect to the average information load. The $\bar{I}_{\text{inf}}$-value is the mode of symmetric distribution of word information load in separate texts and thus determines the most probable word rank in the hypothetical general population of the sample (text corpus), which allows to consider U_R as a usage measure, well suited for the problem of selecting words for educational dictionaries and basic dictionaries of languages (sublanguages) in accordance with their linguistic "importance".

The author is grateful to A.A. Polikarpov for his critical reading of the manuscript and for useful suggestions.

References

Arapov M.V. (1988). *Kvantitativnaja lingvistika*. Moskva: Nauka.

Kondrat'eva V.A. (1972). *Psichologičeskoe obosnovanie putej povyšenija effektivnosti usvoenija leksiki v processe čtenija*. Avtoref. dis. d-ra psichol. nauk. Moskva.

Kromer V.V. (1997a). *Jaderno-veernaja model' vertikal'nogo raspredelenija slov v russkom jazyke.* Novosibirsk: Novosibirskij gos. ped. un-t. Dep. v INION RAN 31.03.97, № 52458.

Kromer V.V. (1997b). Nekotorye osobennosti gorizontal'nogo raspredelenija slov russkogo jazyka po žanrovym gruppam tekstov. In *Razvitie ličnosti v sisteme nepreryvnogo obrazovanija: Tez. dokl. II Meždunar. konf.* : *174–176*. Novosibirsk.

Kromer V.V. (1997c). *Podporno-eksponencial'naja model' general'noj leksičeskoj sovokupnosti anglijskogo jazyka.* Novosibirsk: Novosibirskij gos. ped. un-t. Dep. v INION RAN 18.12.97, № 53134.

Kromer V.V. (1998). *Mera upotrebitel'nosti slova, osnovannaja na psichofizičeskich sootnošenijach.* Novosibirsk: Novosibirskij gos. ped. un-t. Dep. v INION RAN 28.03.98, № 54185.

Kromer V.V. (1999a). Ob odnoj popravke k kanoničeskomu zakonu. In: F.M. Ablaev, K.R. Galilullin (Eds.), *Informacionnye technologii v gumanitarnych naukach: 89–93*. Kazan'.

Kromer V.V. (1999b). *Ocenka nadežnosti častotnych slovarej*. Novosibirsk: Novosibirskij gos. ped. un-t. Dep. v INION RAN 01.07.99, № 54779.

Nešitoj V.V. (1987). Forma predstavlenija rangovych raspredelenij. In: *Učenye zap. Tartus. un-ta 774, 123–134*. Tartu.

Nivre J. (2000). Sparse data and smoothing in statistical part-of-speech tagging. *Journal of Quantitative Linguistic, 7, 1–18.*

Orlov Ju.K. (1978). Statističeskoe modelirovanie rečevych potokov. In: R.G. Piotrovskij (ed.), *Voprosy kibernetiki. Statistika reči i avtomatičeskij analiz teksta. Vyp. 41, 67–95*. Moskva-Leningrad.

Piotrovskij R.G., Bektaev K.B., Piotrovskaja A.A. (1977). *Matematičeskaja lingvistika*. Moskva.

Prochorov A.V. (1982). Logarifmičeski normal'noe raspredelenie. In I.M. Vinogradov (Ed.), *Matematičeskaja enciklopedija: Vol. 3, 408*. Moskva: Sovetskaja enciklopedija.

Prochorov A.M. (Ed.), (1984). Razrešajuščaja sposobnost'. In: *Fizičeskij enciklopedičeskij slovar': 615*. Moskva: Sovetskaja enciklopedija.

Tuldava Ju. (1987). *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki*. Tartu.

Woronczak J. (1967). On an attempt to generalize Mandelbrot's distribution. In: *To Honor Roman Jakobson Vol. 3, 2254-2268*. The Hague: Mouton.

Zasorina L.N. (Ed.) (1977). *Častotnyj slovar' russkogo jazyka*. Moskva: Russkij jazyk.

Glottometrics 5, 2002, 14-21 *To honor G. K. Zipf*

Zipf's Law Everywhere

Wentian Li[1]

Abstract. At the 100th anniversary of the birth of George Kingsley Zipf, one striking fact about the statistical regularity that bears his name, Zipf's law, is that it seems to appear everywhere. We may ask these questions related to the ubiquity of Zipf's law: Is there a rigorous test in fitting real data to Zipf's law? In how many forms does Zipf's law appear? In which fields are the data sets claiming to exhibit Zipf's law?

Keywords: Zipf's law, ranking, language, population, internet, economics, bibliometrics, natural phenomena

1. Testing Zipf's law against alternative functions

Claiming a Zipf's law in a data set seems to be simple enough: if n values, x_i (i =1,2, …,n), are ranked by $x_1 \geq x_2 \geq \ldots x_r \ldots \geq x_n$, Zipf's law states,

$$(1) \qquad x_{(r)} = \frac{C}{r^{\alpha}}$$

where the parameter value, α, is usually close to 1, implies that the $x_{(r)}$ versus r plot on a log-log scale will be a straight line with a negative slope α close to -1. If we assume $x_{(r)}$ as a random variable, from the statistical modeling point of view, Zipf's law is a model of the average of $x_{(r)}$ or log($x_{(r)}$) as a linear function (linear regression) of log(r) (with c = log(C)):

$$(2) \qquad E(\log x_{(r)}) = c - \alpha \log(r).$$

However, visual inspection of the log-log plot of the ranked data is not a rigorous test. What if another functional form fits the same data better? Indeed, there are several functions that have been proposed as alternatives to Zipf's law in fitting the ranked data, such as (i) the Yule distribution (Yule 1925):

$$(3) \qquad x_{(r)} = \frac{C}{r^{\alpha} B^{r}}$$

or in the statistical modeling framework (with c = log(C), b = log(B)):

[1] Address correspondence to: Wentian Li, Center for Genomics and Human Genetics, North Shore LIJ Research Institute, 350 Community Drive, Manhasset, NY 11030, USA. E-mail: wli@nslij-genetics.org

(4) $$E(\log x_{(r)}) = c - \alpha \log(r) - be^{\log(r)};$$

(ii) a variant of the log-normal distribution:

(5) $$E(\log x_{(r)}) = c - \alpha \log(r) - b(\log(r))^2;$$

or, (iii) a variant of the Weibull distribution:

(6) $$E(\log x_{(r)}) = c - \alpha \log(r) - be^{\beta \log(r)};$$

where 0 < β <1.

In all three examples of an alternative function, there is a systematic modulation of the basic power-law structure in Zipf's law. In fact, such systematic deviation from the straight line in log-log plot is indeed present in some claimed Zipf's law patterns (Piqueira et al. 1999), which naturally causes a legitimate concern that some other claimed Zipf's laws in the literature may not be really Zipf's law.

A similar caution was raised in the example of the claimed Zipf's law pattern in DNA oligonucleotide frequencies (Mantegna et al. 1944}. There were many criticisms of this work (see, e.g., Bonhoeffer et al. 1996; Israeloff, Kagalenko, and Chan 1996; Voss 1996; Li 1996). One of the specific criticisms is that the data could be fitted by an alternatively function, the Yule distribution (Martindale and Konopka 1996).

It should be pointed out that it is not enough to reject the Zipf's law only because another function fits the ranked data better. The alternative function should not have too many extra parameters in achieving the better fit. The topic of statistical model selection is extensively discussed in Burnham and Anderson (2002). It is conceivable that we may use either the Bayesian information criterion (BIC) (Schwarz 1976) or Akaike information criterion (AIC) (Akaike 1974; Parzen, Tanabe, and Kitagawa 1998) in selecting Zipf's law among alternatives. Some related ideas were also discussed in Quandt (1964) and Urzua (2000).

2. Two forms of Zipf's law

Besides the familiar form of Zipf's law for ranked data, there is another equivalent form of Zipf's law (Miller 1965). Actually, the second form is the probability density function of $x_{(r)}$, $p(x)$. Considering this simple procedure: switch the rank r and ranked value $x_{(r)}$ axes, then reverse the direction of the $x_{(r)}$. The resulting plot is simply the accumulative distribution (not normalized) of $x_{(r)}$ (see, e.g., Urzua 2000; Rousseau 2002). In mathematical expression, it is:

(7) $$\frac{r(x)}{n} = 1 - \int_{\min(t)}^{x} p(t)dt.$$

Knowing $r(x)$, or equivalently, $x(r)$, the probability density function $p(x)$ can be obtained by

(8) $$p(x) = -\frac{1}{n}\frac{d}{dx}r(x) \quad \text{or,} \quad p(x) = -\left(n\frac{d}{dr}x(r)\right)^{-1}.$$

It can be easily shown that the Zipf's law in Eq. (1) is equivalent to the following form of the probability density function of $x_{(r)}$:

$$(9) \qquad p(x) = \frac{C^{1/\alpha}}{\alpha n} \frac{1}{x^{(1/\alpha)+1}} = \frac{A}{x^{\beta}},$$

with $A = C^{1/\alpha}/n\alpha$ and $\beta = \alpha^{-1} + 1$, Eq. (9) is also a power-law function. The exponent $\alpha = 1$ as proposed originally in Zipf's law leads to $\beta = 2$. Some of the claimed Zipf's law was indeed illustrated as a probability density function (Axtell 2001).

3. Phenomena claiming a Zipf's law pattern

3.1. Word usage in human languages

The variable x is the number of times a word is used in written human languages (Zipf 1932, 1949; Kucera and Francis 1967). The frequency of usage can also be extended to spoken languages (Dahl 1979), non-English or non-Latin languages (Rousseau and Zhang 1992), combination of words (Egghe 2000), etc. Many articles in this volume are devoted to reviews on this example (Rousseau 2002; Altmann 2002; Hřebíček 2002; Montemurro and Zanette 2002).

3.2. City populations

The variable x is the number of people living in a city (Zipf 1949; Hill 1970; Ijiri and Simon 1977; Rosen and Resnick 1980; Gabaix 1999; Knudsen 2001; Soo 2002; Brakman, Garretsen, and Marrewijk 2001). The Zipf's law pattern can be easily checked by obtaining large city population data from a World Almanac, as was done in (Gell-Mann 1994). The city population can also be extended to those of metropolitan area, tribal society, regional areas (Davis and Weinstein 2001), etc. In a recent most extensive analysis of city population in different countries, the exact form of Zipf's law (i.e. $\alpha = 1$) was confirmed in 20 out of 73 countries (Soo 2002).

3.3. Webpage visits and other internet traffic data

In 1997, as a webmaster for a human genetics resource site (http://linkage.rockefeller.edu/), I was curious about whether the number of website visits per month followed the Zipf's law pattern. A quick plot showed it did. Being excited, I wanted to check whether someone else had come up with the same idea before I started to write this up in a publication. My web search ended up at the computer science department of Boston University where the same Zipf's law pattern for webpage visits was already discovered (Cunha, Bestavros, and Crovella 1995)! In the last few years, the study of scaling behaviors in internet traffic (with Zipf's law included) has become one of the hottest topics in applied computer science (Glassman 1994; Crovella and Bestavros 1997; Barford et al. 1999; Huberman et al. 1998; Barabasi and Albert 1999; Breslau et al. 1999; Adamic and Huberman 2002; Mitzenmacher 2003).

3.4. Company sizes and other economic data

This is another example of an easily obtainable data from the World Almanac. A company can be ranked by the number of employees, revenue, profit, market cap, as well as many other measurements. Such ranking can also be done within certain industry or certain geographical locations. The income distribution (Aitchison and Brown 1954; Samuelson 1952; Aoyama et al. 2000) is famously related to Pareto's law (Pareto 1896), which is frequently indistinguishable from the Zipf's law (the only difference being whether the α value is equal to 1 or not). One of the recent large-scale analyses of US company sizes is presented in (Axtell 2001). A constant debate on economic data is whether these are distributed as power-law (e.g. Pareto, Zipf) or as log-normal (Aitchison and Brown 1954; Champernowne 1953; Axtell 2001; Mitzenmacher 2003), or perhaps other distributions (Dagum 1984; Dragulescu and Yakovenko 2001a,b; Azzalini and Kotz 2002}.

3.5. Science citation and other bibliometric data

Similar to the popularity of webpages, popularity of scientific papers can be measured by how many times it is cited by other scientists. Scientists can also be ranked by how many papers he/she publishes (a measure of "productivity"). Other "bibliometric" data include the frequency of library items being loaned/borrowed. A pioneer of bibliometric data analysis was Alfred Lotka (1926). The following papers can be consulted for more details on bibliometric analysis: (Fairthorne 1969; Wyllys 1981; White and McCain 1989; Hertzel 1987; Egghe 1991; Egghe and Rousseau 1990; Osareh 1996a,b; Silagadze 1997; Redner 1998}.

3.6. Scaling in natural and physical phenomena

Since it has been shown that an inverse power-law with exponent α in the ranked data is equivalent to an inverse power-law in the probability density function with the exponent $\beta = (1/\alpha) +1$, and Zipf's law with $\alpha =1$ corresponds to $\beta = 2$, we can bring many more observed scaling behavior (i.e. power-law behavior) (Schroeder 1991) as examples of Zipf's law.

For example, the famous Gutenberg-Richter law states that the number of earthquakes whose magnitude are larger the M is an exponential function of M (Sornette et al. 1996):

$$(10) \quad N(x>M) \propto e^{-bM} \quad \text{with } b \approx 1.$$

Note that Eq.(10) is an accumulative distribution of the probability density function, and earthquake magnitude is a logarithm of the energy released $M \propto \log(E)$. It can be shown that the probability density function for earthquake energy according to Gutenberg-Richter law is $p(E) \propto 1/E^{b+1} = 1/E^2$, same as would be predicted by the Zipf's law.

3.7. Not all data exhibit Zipf's law

Although the title of this article is "Zipf's law everywhere", it is, of course, not literally everywhere. We have already shown examples where systematic deviation is present in the log-log plot of the ranked data (Piqueira et al. 1999; Mantegna et al. 1994). Also, when the size of the data (n) is small, it is usually hard to be convincing that we observe a power-law function. For example, the usage of 20 amino acids in protein sequence does not follow Zipf's law (Gamow and Ycas 1955). The 26 letters also do not follow Zipf's law in an English text.

If the x variable is a derived quantity (as versus a direct observable), the exponent α depends on how x is derived. For example, in a study to rank genes in their ability to classify cancer subtypes (Li and Yang 2002), the (log) likelihood under a statistical discriminant model is used. If this likelihood is normalized by the number of samples in the microarray experiment, the exponent α in the Zipf's plot will be altered. On the other hand, if a more direct measurement is used, it is possible to have a traditional Zipf's law (Furusawa and Kaneko 2003).

4. Conclusions

It is tempting to propose a universal mechanism for Zipf's law because of the impression that Zipf's law is everywhere. Indeed, very general mechanisms were proposed (Yule 1925; Simon 1955), which without doubt would explain a large number of observed Zipf's law patterns (for a review of the explanations of Zipf's law, see, e.g., (Mitzenmacher 2003)).

But is our impression correct? Some of the true Zipf's laws may not be even well known to be a Zipf's law because the data is not presented as a ranked data. As we know the second form of the Zipf's law, we should look for any probability density function of the form $1/x^2$. On the opposite end, many claimed Zipf's law patterns may not be true of Zipf's law after all. Some data might be fitted better by alternative functional forms which nevertheless were not looked into by researchers.

The lesson is that we should pay attention to the data first. We may re-discover new dataset which exhibit Zipf's law, and at the same time, reject some claims of the Zipf's law in the literature. Despite my best efforts to collect all claimed Zipf's law in a webpage (http://linkage.rockefeller.edu/wli/zipf/), such efforts seem to be less than perfect, and there are always false claims and missing ones.

Acknowledgements

I would like to thank Jeff Robbins for proofreading the first draft of the paper.

References

Adamic, L.A., Huberman, B.A. (2002). Zipf's law and the internet. *Glottometrics 3, 143-150.*

Aitchison, J., Brown, J.A.C. (1954). On criteria for descriptions of income distribution. *Metroeconomica 6, 88-98.*

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19, 716-723.*

Altmann, G. (2002). Zipfian linguistics. *Glottometrics 3, 19-26.*

Aoyama, H., Souma, W., Nagahara, Y., Okazaki, M.P., Takayasu, H., Takayasu, M. (2000). Pareto's law for income of individuals and debt of bankrupt companies. *Fractals 8, 293-300.*

Axtell, R.L. (2001). Zipf distribution of U.S. firm sizes. *Science 293, 1818-1820.*

Azzalini, A., Kotz, S. (2002). *Log-skew-normal and log-skew-t distributions as models for family income data*. University of Padua Department of Statistical Sciences, preprint.

Barford, P., Bestavros, A., Bradley, A., Crovella, M. (1999). Changes in web client access patterns: characteristics and caching implications. *World Wide Web 2, 15-28.*

Barabasi, A.L., Albert, R. (1999). Emergence of scaling in random networks. *Science 286, 509-512.*

Bonhoeffer, S., Herz, A.V.M., Boerlijst, M.C., Nee, S., Nowak, M.A. May, R.M. (1996), Explaining 'linguistic features' of noncoding DNA. *Science 271(5245), 14-15.*

Brakman, S., Garretsen, H., Marrewijk, C. van (2001). *An Introduction to Geographical Economics*. Cambridge, England: Cambridge University Press.

Breslau, L., Cao, P., Fan, L., Philips, G., Shenker, S. (1999). Web caching and Zipf-like distributions: evidence and implications. *Proceedings of IEEE Infocom'99, 126-134.*

Burnham, K.P., Anderson, D.R. (2002^2). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Berlin: Springer-Verlag.

Champernowne, D. (1953). A model of income distribution. *Economic Journal 63, 318-351.*

Crovella, M., Bestavros, A. (1997). Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking 5, 835-846.*

Cunha, C.R., Bestavros, A., Crovella, M.E. (1995). Characteristics of WWW client-based traces. Boston University Computer Science Department, *Technical report TR-95-010.*

Dagum, C. (1984). Income distributions models. In: Kotz, S. Johnson, N.L., Read, C.B. (eds.*), Encyclopedia of Statistical Sciences, vol. IV: 21-34*. New York: Wiley.

Dahl, G. (1979). *Word Frequencies of Spoken American English*. Essex, CT: Verbatim.

Davis, D.R., Weinstein, D.E. (2001). *Bones, bombs and break points: the geography of economic activity.* National Bureau of Economic Research, working paper 8517.

Dragulescu, A., Yakovenko, V.M. (2001a). Evidence for the exponential distribution of income in the USA. *The European Physical Journal B, 20, 585-589.*

Dragulescu, A., Yakovenko, V.M. (2001b). Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A, 299, 213-221.*

Egghe, L. (1991). The exact place of Zipf's and Pareto's law amongst the classical informetric laws. *Scientometrics 20, 93-106.*

Egghe, L. (2000). The distribution of N-grams. *Scientometrics 47, 237-252.*

Egghe, L., Rousseau, R. (1990). *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Amsterdam-New York: Elsevier.

Fairthorne, R.A. (1969). Empirical hyperbolic distributions (Bradford Zipf Mandelbrot) for bibliometric description and prediction. *Journal of Documentation 25, 319-343.*

Furusawa, C., Kaneko, K. (2003). Zipf's law in gene expression. *Physical Review Letters 90, 88-102.*

Gabaix, X. (1999). Zipf's law for cities: an explanation. *Quarterly Journal of Economics, 114, 739-767.*

Gamow, G., Ycas, M. (1955). Statistical correlation of protein and ribonucleic acid composition. *Proceedings of the National Academy of Sciences 41(12), 1011-1019.*

Gell-Mann, M. (1994). *The Quark and the Jaguar*. New York: Freeman.

Glassman, S. (1994). A caching relay for the world wide web. *Computer Networks and ISDN Systems 27(2), 165-173.*

Hertzel, D.H. (1987). Bibliometrics, history of the development of ideas. In: *Encyclopedia of Library and Information Science vol. 42, suppl. 7, 144-211*. New York: Dekker.

Hill, B.M. (1970). Zipf's law and prior distributions for the composition of a population. *Journal of the American Statistical Association 65, 1220-1232.*

Hřebíček, L. (2002). Zipf's law and text. *Glottometrics 3, 27-38.*

Huberman, B.H., Pirollo, P.L.T., Pitkow, J.E., Lukose, R.M. (1998). Strong regularities in world wide web surfing. *Science 280, 95-97.*

Ijiri, Y., Simon, H.A. (1977). *Skew Distributions and the Sizes of Firms*. Amsterdam: North-Holland.

Israeloff, N.E., Kagalenko, M., Chan, K. (1996). Can Zipf distinguish language from noise in noncoding DNA? *(letters), Physical Review Letters 76(11), 1976.*

Knudsen, T. (2001). Zipf's law for cities and beyond – the case of Denmark. American *Journal of Economics and Sociology 60, 123-146.*

Kucera, H., Francis, W.N. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Li, W. (1996). Comments on 'Bell curves and monkey languages' (letters). *Complexity 1(6), 6.*

Li, W., Yang, Y. (2002). Zipf's law in importance of genes for cancer classification using microarray data. *Journal of Theoretical Biology 219(4), 539-551.*

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16, 317-323.*

Mantegna, R.N., Buldyrev, S.V., AL Goldberger, A.L., Havlin, S., Peng, C.P., Simon, M., Stanley, H.E. (1994). Linguistic features of noncoding DNA sequences. *Physical Review Letters, 73, 3169-3172.*

Martindale, C., Konopka, A.K. (1996). Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers & Chemistry 20, 35-38.*

Miller, G.A. (1965). Introduction. In: *The Psycho-Biology of Language*. Cambridge, MA: MIT Press.

Mitzenmacher, M. (2003). A brief history of generative models for power law and lognormal distributions. Harvard University EECS Department, preprint.

Montemurro, M.A., Zanette, D.H. (2002). New perspectives on Zipf's law in linguistics: from single texts to large corpora. *Glottometrics 4, 87-99.*

Osareh, F. (1996a). Bibliometrics, citation analysis and co-citation analysis: a review of literature I. *Libri 46, 149-158.*

Osareh, F. (1996b). Bibliometrics, citation analysis and co-citation analysis: a review of literature II: *Libri 46, 217-225.*

Pareto, V. (1896). *Cours d'Economie Politique*. Geneva: Droz.

Parzen, E., Tanabe, K., Kitagawa, G. (1998). *Selected Papers of Hirotugu Akaike*. Berlin: Springer.

Piqueira, J.R., Monteiro, L.H., Magalhaes, T.M. de, Ramos, R.T., Sassi, R.B., Cruz, E.G. (1999). Zipf's law organizes a psychiatric ward. *Journal of Theoretical Biology 198, 439-443.*

Quandt, R.E. (1964). Statistical discrimination among alternative hypotheses and some economic regularities. *Journal of Regional Science 5, 1-23.*

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B 4,131-134.*

Rosen, K.T., Resnick, M. (1980). The size distribution of cities: an examination of the Pareto law and primacy. *Journal of Urban Economics 8, 165-186.*

Rousseau, R., Zhang, Q. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics 24(2), 201-220.*

Rousseau, R. (2002). George Kingsley Zipf: life, ideas, his law and informetrics. *Glottometrics 3, 11-18.*

Samuelson, P.A. (1992). Graduated income taxation, which reduces inequality, leaves Pareto's coefficient invariant: a pseudo-paradox that debunks Pareto's coefficient. *Journal of Economic Perspectives 6, 205-206.*

Schroeder, M. (1991). *Fractals, Chaos, Power Laws*. New York: Freeman.

Schwarz, G. (1976). Estimating the dimension of a model. *Annals of Statistics 6, 461-464.*

Silagadze, Z.K. (1997). Citations and the Zipf-Mandelbrot's law. *Complex Systems 11, 487-499.*

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika 42, 425-440.*

Soo, K.T. (2002). Zipf's law for cities: a cross country investigation. London School of Economics, preprint.

Sornette, D., Knopoff, L., Kagan, Y.Y., Vanneste, C. (1996). Rank-ordering statistics of extreme events: application to the distribution of large earthquakes. *Journal of Geophysical Research 101, 13883-13893.*

Urzua, C.M. (2000). A simple and efficient test for Zipf's law. *Economics Letters 66, 257-260.*

Venables, W.N., Ripley, B.D. (1999^3). *Modern Applied Statistics with S-PLUS*. Berlin: Springer.

Voss, R.F. (1996). Linguistic features of noncoding DNA sequences – Comment" (letters), *Physical Review Letters 76(11), 1978.*

White, H., McCain, K.W. (1989). Bibliometrics. *Annual Review of Information Science Technology* 24,119-186.

Wyllys, R.E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends 30, 53-64.*

Yule, G.U. (1925). A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F.R.S. *Philosophical Transactions B 213, 21-87.*

Zipf, G.K. (1932). Selected Studies of the Principle of Relative Frequency in Language. Cambridge, MA: Harvard University Press.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.

Glottometrics 5, 2002, 22-28 *To honor G.K. Zipf*

Zipf's Tool Analogy and Word Order

Gertraud Fenk-Oczlon[1]
August Fenk

> *As we proceed from the artisan down the bench we shall proceed from (a) the ever smaller, lighter, and more frequently used tools to (b) the ever larger, heavier, and less frequently used tools.*
> *G. K. Zipf, 1949: 62*

Abstract. This article starts with Zipf's (1949) "Tool Analogy", where the artisan arranges and re-designs his tools in a way minimizing his total work; as a result, more frequently used tools tend to be nearer to him (better accessible), smaller and multifunctional. We then argue that short distance, small size and multifunctionality reflect not only a high overall relative frequency of usage, but in particular a high frequency of usage in the first steps of a variety of complex working procedures. Tool order – word order? This extended Tool Analogy fits to the tendency of more frequent words to obtain initial positions in frozen binomials (Fenk-Oczlon 1989) and the new finding (Fenk & Fenk-Oczlon 2002a,b) that the short, frequent and multifunctional function words tend to concentrate in the first part of sentences.

Keywords: Zipf's Tool Analogy, word frequency, word order, freezes, function words, cognitive economy, information theory

1. Zipf's Tool Analogy

Chapter Three in Zipf (1949) starts with the explication of what he calls "Tool Analogy“ – "tools" in analogy e.g. to verbal expressions such as words. It is the aim of the present study to investigate if the arrangement of tools in Zipf's analogy corresponds to the arrangement of words in phrasal conjuncts and in sentences: Tool order – word order?

Before going on to some general remarks on the use of such analogies introduced into scientific communication and before investigating the potential of the Tool Analogy as an intelligent illustration of concrete empirical phenomena within the domain of linguistics, let us give a short characterization of this analogy in the words of Zipf (1949):

An artisan "must survive by performing certain jobs for us with his tools as economically as possible. Beyond that we do not care. Thus we do not care how many tools he uses, nor how he alters their size, shape, weight, and usage, nor how he arranges them on the board, as long as he performs the specific fixed job with a minimum of total work" (p.58). This total work is the product of $f \times m \times d$ (f = frequency of usage of a certain tool, m = the mass or size of the tool, and d = the distance "of a given tool to be its round-trip distance to the artisan's lap and return…" p. 59). "However since the artisan is obliged to use his tools with a

[1] Address correspondence to: Gertraud Fenk-Oczlon, Department of Linguistics and Computational Linguistics, University of Klagenfurt, Universitaetsstrasse 65-67, A-9020 Klagenfurt. E-mail: Gertraud.Fenk@uni-klu.ac.at

maximum economy *he must arrange the n tools of his shop in such a way that the sum of all the products of* $f \times m \times d$ *for each of the n tools will be a minimum*" (p.59). It is the question of "Close Packing" which is important, because "'close packing' will at all times decrease the *d* distance of the tools and thereby decrease the work of using them, regardless of the size or mass of the tools in the shop" (p.60). Furthermore, "there is an economy in a small size" of tools (p.60).

> "Therefore the magnitude of the Force of Abbreviation will tend to decrease in direct proportion to the distance of the tool from the artisan; the farther that a given tool is from the artisan, the proportionately less the comparative economy will be in reducing its size by a given amount. *Hence in redesigning his tools the artisan will lay a premium upon the reduction of the sizes of all tools in proportion to their nearness to him.*
>
> As a result of the above, we may expect to find in our artisan's shop, as a consequence of years of redesigning, that there will be a tendency for the sizes of tools to stand in an inverse relationship to their nearness to the artisan (i.e., the nearer tools will be the smaller). We shall henceforth call this inverse relationship between size and nearness the Principle of the Abbreviation of Size" (Zipf 1949:61).

> "Furthermore, as the frequency of the easiest tool increases (while its mass decreases), the ever nearer to the artisan the tool will be moved because of the exigencies of the 'minimum equation'; and the ever nearer to the artisan that the tool is moved the ever greater will be the Force of Abbreviation in reducing its size" (Zipf 1949:62).

2. On the potential of the Tool Analogy as a cognitive-communicative tool

A general view on language as an "organon", or as a "mental organ" or as a cognitive-communicative "tool" is neither new nor very concrete in detail. Much more convincing is the tool character of the specific analogy introduced by Zipf and the tool character of the spatial metaphors used by Zipf when explicating his analogy. The tools being "nearer" to the artisan are such a metaphor. Powers (1998:152) identifies the term "distance" in Zipf's analogy with "access time".

Seemingly, the potential of Zipf's analogy has not been exhausted so far. Rather recent findings in quantitative linguistics can be illustrated or "explained" by this analogy – at least if we extend it in a certain respect or make it more explicit in this certain respect:

Most jobs to be done by our artisan – a shoemaker, a potter, a coppersmith – require not only a single tool, but a series of tools in a non-arbitrary order. Usually these series will start with rather common and unspecific tools of the handicraft in question before proceeding to more specified tools. Thus, the more common tools obtain not only a high overall frequency of usage, but especially a high frequency of usage in the first and basic steps or operations of a wide range of complex procedures. Special requirements and special tools, e.g. for different decorations of the product, follow – if at all – later in this procedure. Both of the following linguistic findings can be "explained" by the tool analogy, if it is extended or more explicit in this aspect, and both of them are special cases of the rule "the more frequent before the less frequent".

3. Frequency as a determinant of order: more frequent words tend to be placed before less frequent words!

3.1. The tendency of the more frequent word to obtain the initial position in frozen binomials (freezes)

More frequently used words are easier for the speaker to call up and more expectable for the hearer. In order to achieve a constant flow of linguistic information and to avoid peaks of information, such informationally poorer elements should be placed at those positions which are *per se* associated with higher informational content. This was the main argument for the hypothesis (Fenk-Oczlon 1989) that in freezes, i.e. frozen conjoined expressions or binomials, such as *knife and fork, peak and valley, salt and pepper*, the more frequent word would tend to obtain the initial position. The predictive power of this new rule was tested on the basis of 400 freezes from English, Russian and German and was compared to the predictive power of rules previously proposed by other authors, such as “short before long”, “the first word has fewer initial consonants than the second”, “front vowel before back vowel”, and “semantic principles” (such as the me-first principle). Token frequencies of the single words constituting the frozen binomials were taken from Thorndike & Lorge, Josselson, Meier, Ruoff. The result: With 84% of the predictions being correct (i.e. 337 of 400 freezes) the new rule achieved by far the highest accuracy.

In the context of the present paper, the most interesting rival in this competition was the rule “short word before long word”. (As we all know e.g. from Zipf’s work, there is a strong inverse relationship between frequency of usage and length of the respective words.) The result of the direct comparison: "High frequency before low frequency” scored with 337 hits, the rule “short before long” with only 152 hits. (To some degree, this difference is a result of a handicap of the latter rule; it is not applicable in cases of equal number of syllables of first and last word). In 145 of these 152 freezes where the word order can be “explained” by “short before long”, the order can as well be explained by “more frequent before less frequent”. This enormous overlap, together with the higher rate of hits of the frequency rule, is one of several arguments saying that the frequency rule represents a principle that is superordinate to the competing rules.

In his “Tool Analogy” Zipf characterizes the “dynamic” interrelationship between “ease” (small product of $m \times d$) and frequency as follows: “In short, *greater frequency makes for greater ease which makes for greater frequency and so on*” (Zipf 1949:62). This formulation suggests that in this process of a mutual build up between increase of frequency and increase of “ease” the initial impulse usually will come from the variable “frequency”.

3.2. The tendency of function words to concentrate in the first part of sentences[2]

In an experimental study by Auer, Bacik & Fenk (2001) on the memory for sentences a text of Glasersfeld (1998) was presented auditorily. A tone at the end of some of the sentences (n = 10) signalled to the subjects that they should try to immediately recall as many words as possible from this sentence.

Reanalyzing the data in order to investigate word-class specific effects on recall (Fenk & Fenk-Oczlon 2002a) we made the following observation: In the sentences presented, function words (such as pronouns, articles, conjunctions,…) dominantly occurred in the first quarter of the sentences, whereas content word (nouns, verbs, adjectives, adverbs) did so rather in the

[2] So far, the new empirical findings summarized in this section have only been “published” in conference papers (Fenk & Fenk-Oczlon 2002a, b); a full version will follow (Fenk & Fenk-Oczlon, in preparation).

last (see Figure 1, left panel). This was a problem for the statistical evaluation of the recall scores – recall scores in absolute terms were meaningless and had to be related to the proportion of function words and content words occurring in the relevant portion of the sentence – but interesting from the point of view of quantitative linguistics. Were the differences found in the within-sentence distribution of function words and content words a specific characteristic of this one author Glasersfeld or a rather general regularity?

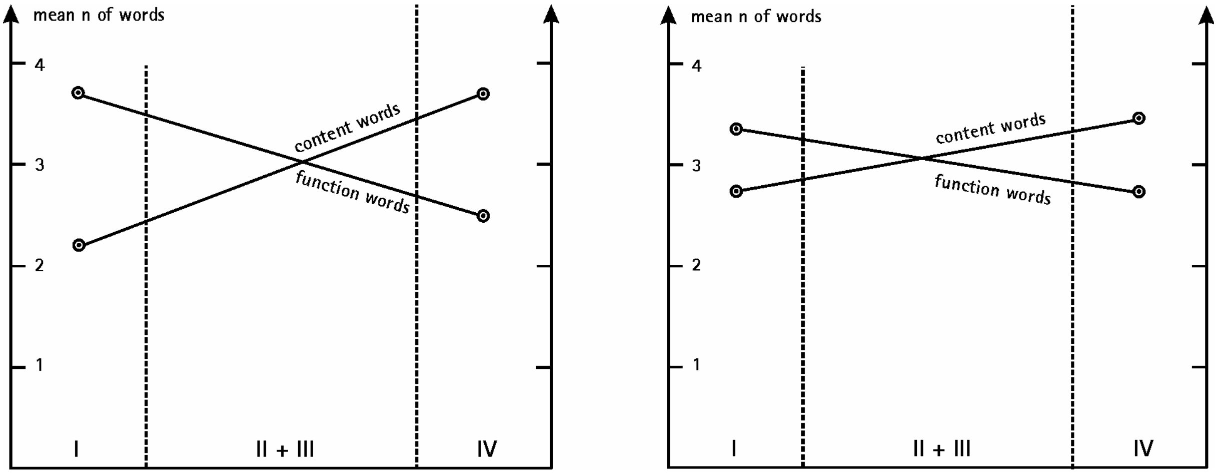


Figure 1: Mean number of function words and content words in the first quarter (I) and in the last quarter (IV) of sentences
Left panel: Mean of 10 sentences from the author Glasersfeld (1998)
Right panel: Mean of 100 sentences (10 sentences from each of 10 authors)

In order to find a first answer to this question German texts of 9 additional authors were analysed: 10 sentences (each third sentence of a text, where possible) from each of 4 scientific texts and 5 literary texts. Results are illustrated in Figure 1, right panel. Table 1 and Table 2 present the numerical values and the significance of results. These results suggest that the word-class specific within-sentence distribution can be generalized for contemporary German texts.

Table 1
Mean frequency of function words and content words in the first quarter (I) versus last quarter (IV) of 100 sentences (10 sentences from each of 10 authors)

	I		IV	differences
function words	3.36	>	2.67	significant, p<1%
content words	2.74	<	3.46	significant, p<1%

Table 2
Mean frequency of function words versus content words within the first quarter (I) and the last quarter (IV) of 100 sentences (10 sentences from each of 10 authors)

	function words		content words	differences
quarter I	3.36	>	2.74	significant, p<5%
quarter IV	2.67	<	3.46	significant, p<1%

In connected discourse many sentences will refer to what was mentioned in the preceding sentences. This reference will most commonly occur in the first part of the sentence ("Thema" before "Rhema", topic before comment, old before new), and function words might play a dominant role in this sort of reference.

If this is an appropriate explanation for the regularity found, this regularity should not be restricted to German texts, but should rather be universal in its essential respect, i.e. the decrease of function words and increase of content words while the sentence proceeds (Table 1). The starting points of decrease and increase will, however, vary from language to language, dependent, for instance, on the proportion of function words in the specific language. All three patterns shown in Figure 2 seem to be possible. The panel in the middle of Figure 2 represents the proportions found in German, and the analysis of Müller (in progress) indicates that the left panel might represent a pattern characterizing the proportions in Roman languages.

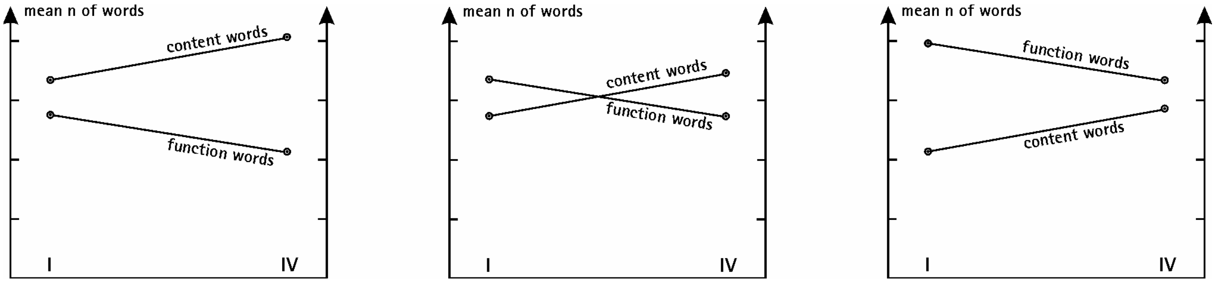


Figure 2. Three different patterns resulting from a within-sentence decrease of function words and increase of content words from changing starting points.

From our regularity – the within-sentence decrease of function words – we may derive a further regularity (Fenk & Fenk-Oczlon 2002b). In those languages, where our first regularity applies, the following regularity regarding the distribution of different word lengths within sentences will apply as well: the mean length of words will increase from the beginning to the end of sentences. The reason for this is that function words are not only extremely frequent, but also – for exactly this reason, as we know e.g. from Zipf (1929, 1949) – relatively short. The prevailing of the (very frequent and therefore) rather short function words in the first part of sentences might contribute to or even account for Behaghel's (1909) "Gesetz der wachsenden Glieder". It would be an interesting attempt to study these regularities in bigger and machine-readable text corpora.

Postponing "heavy" and "new constituents" does not only facilitate comprehension. Arnold et al. (2000: 28) put stress on the fact that it (also) facilitates "processes of planning and production". We might add that this applies not only to the activities of speakers, but the planning and production processes of "other" artisans as well. Anyway: Since we are incessantly endeavoured to anticipate how speech will continue, when we are in the role of the hearer, and since, when in the role of the producer, we are always also the hearer and controller of our own production, producing and listening will follow very similar strategies. "Ease" in active planning and production will correspond to "ease" in anticipation and comprehension.

4. Concluding remarks

It is a fascinating attempt to explore the heuristic potential and the implications of spatial analogies and spatial metaphors. The use as well as the risks of these heuristic tools – be it Zipf's Tool Analogy or one of the more actual neural network analogies – lies in their graphic quality. We have tried to describe an additional parallel between the tool repertoire of the artisan and our lexical repertoire: Tool order – word order!

In Zipf's analogy (Zipf 1949:59) the "relative frequency" of usage is a central term. In cognitive psychology, the learning of relative frequencies, or the "sensitivity" to relative frequencies is also seen as a fundamental mechanism; without this mechanism we would not even be able to identify word units within the speech stream (Zacks & Hasher 2002:28f). Distributions of relative frequencies and probabilities are, moreover, the central topic of the "formal theory of communication", i.e. of the information theory. This theory lacks the concreteness of Zipf's Tool Analogy. To say it positively: it does not need any demons or any "flesh-and-blood artisan" and allows for a more general, integrative and quantitative description of relevant phenomena. In terms of this theory we may say: a higher frequency of the usage of elements goes hand in hand with both, their reduction in "size" (phonological complexity, duration) and informational content ("higher familiarity", "higher accessibility", "higher availability", "lower cognitive costs", "easier to process"). This means: less time for communicating less information! Thus, these reduction processes provide a "constant" and economic flow of linguistic information. And so does the tendency to localize those elements in the initial positions which carry – overall and/or in the specific context – a lower amount of information: the "topic" which comes before "comment", the "old" that comes before the "new", carries low information in this context and – by ways of transitional probabilities (redundancy) – lowers the information of what follows – the "comment", the "new", or the second partner in a binomial.

All the statistical laws discussed above (short before long, more frequent before less frequent) seem to contribute to an efficient communication by contributing to the principle or "covering law" of a constant flow of linguistic information. The central units of this rhythmically organized information flow are clauses with a relatively "constant" duration containing a relatively "constant" number of elements (syllables) and a relatively "constant" amount of information (Fenk-Oczlon & Fenk 2002: 224): Are these units, then, "packages" with an optimal size for cognitive "handling"? This aspect of "Close Packing" (Zipf 1949: 60) would be worthy of a separate study entitled "Zipf's Tool Analogy and the optimal size of clauses".

References

Arnold, J.E., Wasow, Th., Losongco, A. & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language 76, 1, 28–55.*

Auer, L., Bacik, I. & Fenk, A. (2001). Die serielle Positionskurve beim Behalten echter Sätze. *Paper presented at the 29. Österreichische Linguistiktagung, October 26-27 in Klagenfurt.*

Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen 25, 110-142.*

Fenk, A. & Fenk-Oczlon, G. (2002a). The decay of function words in the recall of sentences of different size. Paper presented at "Wortlängen in Texten. Internationales Symposium zur quantitativen Textanalyse". June 21-23 in Graz/Seggau.

Abstracts, http://www-gewi.uni-graz.at/quanta/programm.htm

Fenk, A. & Fenk-Oczlon, G. (2002b). Funktions- und Inhaltswörter in der statistischen Binnenstruktur von Sätzen. Paper presented at the 30. Österreichische Linguistiktagung, December 6-8 in Innsbruck.
Abstracts, http://www.uibk.ac.at/c/c6/c604/abstract.html

Fenk, A. & Fenk-Oczlon, G. (in press). Within-sentence distribution and retention of content words and function words. In P. Grzybek (ed.) *Word length studies and related issues.*

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics 27*, *517–556.*

Fenk-Oczlon, G. & Fenk, A. (2002). The clausal structure of linguistic and pre-linguistic behavior. In T. Givón & B. F. Malle (eds.) *The evolution of language out of pre-language: 215-229.* Amsterdam: John Benjamins Publishing Company.

Glasersfeld, E. von (1998). Konstruktivismus statt Erkenntnistheorie. In W. Dörfler & J. Mitterer (eds.*) Ernst von Glasersfeld – Konstruktivismus statt Erkenntnistheorie: 11-39.* Klagenfurt/Celovec: Drava Verlag.

Müller, B. (in preparation). *Die statistische Verteilung von Wortlängen und Wortklassen in lateinischen und italienischen Sätzen*. Phil. Diss., University of Klagenfurt.

Powers, D.M.W. (1998). Applications and explanations of Zipf's law. In: D.M.W. Powers (ed.): *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning: 151-60.* ACL.

Zacks, R. T. & Hasher, L. (2002). Frequency processing: a twenty-five year perspective. In: P. Sedlmeier & T. Betsch (eds.) *etc. frequency processing and cognition: 21-36.* Oxford: Oxford University Press.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology 40*, 1-95.

Zipf, G. K. (1949). *Human behavior and the principle of least effort. An introduction to human ecology*. Cambridge, Mass.: Addison-Wesley [2nd ed. New York: Hafner 1972].

Glottometrics 5, 2002, 29-50 *To honor G. K. Zipf*

The Unexpected Fundamental Influence of Mathematics upon Language

Wolfgang Hilberg[1]

Abstract. The functional structure of human language networks in the brain could be revealed in an indirect way by measurements in the abstraction level of words. The result is a very large deterministic graph or network, respectively, which was unknown in mathematics up to now. The whole network can only be represented in a matrix. Following Shannon's theory, it displays optimum properties for information processing (maximum entropy). The structure of the network can be subdivided by introducing word classes with increasing magnitudes which could contribute to an understanding of the biological generation of networks. The hypothesis is that such facts are valid for all natural languages. Differences will exist only in the individual distribution of matrix dots. That means, speaking precisely, that every language has a distinct individual network structure of its own. Surprisingly it can be shown that the superior general type of the universal network structure can be generated by statistical experiments. The properties of these networks were compared with those of natural language networks which are definitely not statistical. Finally the enigma of the famous Zipf-diagram can be disclosed by observing networks and text paths which run inside of them along existing connections from node to node. A staircase curve emerges, which is a better description of reality than a smoothed power law. All this can be repeated by experiments, which means that eventually we found a transition from descriptive to constructive science. Therefore the new ideas could be applied immediately also in technology.

Certainly the basic biological language structure arose a long time ago. Later on the typical patterns of the network connections for different language families should have evolved separately and were almost certainly accompanied by optimization processes for maximum entropy. Nowadays the details of the connection patterns for any language have to be learned anew by every child, and in this process, unusual alterations are not allowed by its language community.

Keywords: text processing, language network, hierarchical network system, word classes, multistage maximum entropy, Zipf´s law, association matrix, matrix dots for connections, functional network structure

1. Language and randomness – an antagonism

Natural language is such an overwhelmingly extensive, precise, orderly, complicated, and nevertheless flexible world, that one could imagine only a great number of geniuses have created such a wonderful system. The opposite, however, could be called "chaos". It is apparently not controlled by the ideas of geniuses but by arbitrariness and randomness. The order of language and the disorder of chaos are, at first glance, incompatible. For example it is certain by all means that when we speak or write some sentences we do not choose the words

[1] Address correspondence to: Wolfgang Hilberg, TU Darmstadt, Fachbereich 18, Merckstr. 25, D-64283 Darmstadt. E-mail: hil@dtro.tu-darmstadt.de

depending on probabilities. We know immediately and exactly whether the word sequence is correct or incorrect. And yet, when languages were created, randomness and its rules must have played an important part (as will be shown in this paper as well). Suspicions about this already existed for a long time: It is revealing that the mathematician Benoit Mandelbrot, later on the inventor of the theory of chaos, struggled hard to understand language and its enigmas (Mandelbrot 1953), unfortunately with limited success. The bitter dicussions which continued for many years between Mandelbrot and another mathematician, J.A. Simon, are not forgotten in the scientific community. Simon doubted the validity of Mandelbrot's theories (Mandelbrot 1953, 1959, 1961a,b; Simon 1955, 1960, 1961a,b, 1963; Rapoport 1982; Li 1992).

2. What can be said about language?

Let us risk a second look and consider popular agreements which widely exist about language. Who, for example, would oppose the following, almost trivial statement: "The creation of natural languages is the greatest achievement of man". In this respect one can refer to the early philosophers, who stated that thinking is only possible through language (this was the first definition of thinking). Of course no philosopher could be understood without language. Wittgenstein (1989, 2001), a modern philosopher, sums up the situation with the statement: "philosophy is only language criticism".

Natural language text is a coherent sequence of elements that are called words. Especially printed text, which is the object here, is restricted to a general meaning of words and does not contain further information which could also be included in handwritten or spoken words. Words can only be arranged in a text in a special manner. For example, if we would cut out the words of a text in a book, and then whirl them around like in a lottery, and if we would connect them again in a random sequence, a completely senseless text would result. Let us note here that the vocabulary and the frequency of words would remain unchanged. Language has to obey necessary rules, which are referred to as "grammar". These rules are very strict, as we can experience when we learn a foreign language. It often depends on the position of a single word. For example, if we cut out a word of a text at an arbitrary position and replace it randomly by another word, everybody will immediately perceive this change. We are often already startled by the mere interchange of the sequence of two immediately following words.

What is behind this? Basically, we could say that it is the brain that produces language or text, respectively. Therefore, the brain has to be in command of the grammar of language. Within a given text, the effect of this grammar can be recognized, i.e. the properties of the brain-product "text" are apparently not purely accidental. In certain respects, words in the text-sequence are even predictable, which helps when we have to understand other people. Very strong linguistic agreements exist within a language community, e.g. in the German language.

The unique properties of this system, called "language", can be conceived by looking at the modern processing of text by computers, for example, at the task of translating text precisely from a first language into a second. Many decades ago, it was a great surprise for researchers that this task was so very extensive and complex. The system "language" consists not only of many hundred thousands of different words (vocabulary) but also of a never-ending number of grammatical rules: Any exception from well-known rules has to be interpreted as a special rule – and everybody knows the large number of exceptions. Considering these facts it is understandable that in Artificial Intelligence (AI), in spite of unbelievably efficient computer technology, researchers failed up to this day to construct a language computer which is as effective as an intelligent human adult – even when restrictions were made to the size of vocabulary and the number of grammar rules.

Besides this modern attempt of utilizing computers by AI-scientists (key word: computer linguistics), mainly inspired by the ideas of Noam Chomsky, who looked with fascination at the endless problems of grammar, there also exists another attempt, which is called "connectionism". Here, to begin with, one has to forget that grammar exists. Personal experience shows that this is not very easy. Instead of grammar one has to observe the structure of a biological system, which is able to produce language. Everybody knows the data: In our brain we have a huge network consisting of about 10^{11} neurons and a much larger number of nerve fibre connections, called axons and dendrites (about 10^4 per neuron i.e. on the whole about 10^{15} connections). But what is the structure of this network? Surgical incisions in the brain have shown that there are unbelievably dense interconnections, a three dimensional texture, throughout this huge complex microbiological network, so that we surely will never have the possibility of directly observing and extracting the system structure.

3. The indirect approach for determining functional brain structure

The connectionistic approach starts out with looking at the single biological neuron and its computer simulation. Many of such so-called artificial neurons can be connected to realize technical systems. Being unaware of any natural model of the connection patterns, suitable network structures were simply invented, founded on technical and mathematical criteria. These artificial neural networks had indeed a certain efficiency in practice when utilized for classification problems (system examples are: Perceptron, Learning Matrix, Hopfield Network, etc.). Problems of higher complexity, however, e.g. understanding or generating text, could not be solved. In order to proceed, we should make a new approach. It is not enough to learn from nature the essential features of a single neuron, we should also learn what type of functional network can be found in nature for the connection of neurons – networks which are in fact capable of solving real complex problems. This will produce surprising results.

Recent investigations have shown (Hilberg 1997a, 1999, 2000a,b, 2002) that we can reveal in an indirect way parts of the functional network structure of the language brain. The idea was that it should be possible to draw conclusions from system output signals, i.e. text, to the inner system structure, which is the generator, called "brain". That is especially possible when the generating system consists of a large number of similar nodes (neurons) and a much larger number of similar connections (nerve fibres).

The working hypothesis: We enter this complex brain system not at the low level of neurons with its weights and thresholds and multiple inputs and outputs but at an easily accessible higher level of abstraction, which is the level of words of natural language (basis level). Each of these words may be stored in a group of some neurons whose detailed properties are neglected (this neglection represents in a literal sense the "abstraction"). Finally the nodes remain as abstracted results having only a place, connections and perhaps a code-name. (In the other direction, to low levels, where the preprocessing of words takes place we should have to consider however the detailed properties of neurons which is the requirement for usual artificial neural networks. But that is not the object here). Words in text are understood as elements, which have existence and meaning irrespective of any particular spelling or pronunciation, that is, words should be considered as the elements in printed text. Each of these words are stored locally in the brain, an assumption which is suggested by brain physiology. (The rival concept of holographic storage is physically different but functionally equivalent). It can be supposed that in every node at least one neuron per word is needed, whose numerous axons and dendrites may control signals both to nodes in higher level networks (Hilberg 1997a, 2000a) and to motor units at lower levels, which are necessary for speaking and writing. Specific sensor units for reading and hearing should be present in lower levels as well. That is, many connections should exist both between neurons in a network and between neurons of different networks. In respect of existing biological association processes: Following the natural model it would be necessary to introduce some thousand additional connections per node for "wiring" the necessary association processes. For the realization of mathematical simulations however it is more convenient not to simulate a vast number of connections but to assume a locally stored technical code in each network node and to use the well known principles of technical associative memories.

After the discussion of the premises above, the following steps in the research will be logical. At first the investigation was restricted to the immediate neighbourhood (juxtaposition) of words in text. Later on a more distant neighbourhood was considered by systematically proceeding into higher levels of abstraction above the level of words (Hilberg 1997a, 2000a). However that was a greater venture and cannot be the object of this paper. Therefore in the next chapters, dealing here with the basic word level only, we are not allowed to pay attention to a context farther than the immediately following next word. Words are interpreted as contents of nodes of the functional network in the layer of the basis level (not to be confounded with a physiological layer). Fig. 1 shows in principle how text is generated in the functional network by activating sequentially one node after the other in a so-called "text path" (note: any node can be in two states, i.e. active or passive, something that can be accomplished by a single neuron as was mentioned before). The sequence of these words can be delivered by converters to an output as a running text. Conversely when hearing or viewing text from outside, signals from preprocessing circuits enter the basis level and activate the sequence of word nodes in a text path.

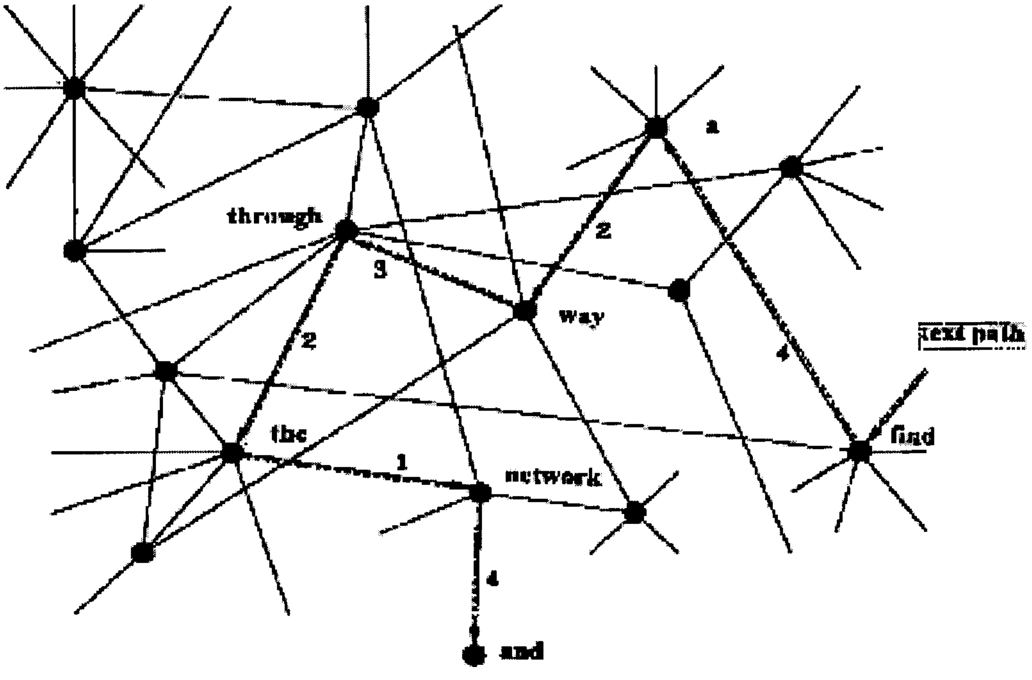


Fig. 1. Principle of a network with nodes representing stored words and connections between the nodes, so that a path from node to node corresponds with a natural language text

4. The measurements

The local storage of words in nodes and their embedding in a functional network by means of directional connections between them is an attractive hypothesis. The model is very simple and yet all possibilities of natural language concerning immediate neighbourhood are covered, provided that a large vocabulary and a sufficiently large volume of learning text was used (again one should take into consideration that the neglected far-reaching context influence has to be considered elsewhere by additional networks and methods, as could be shown in Hilberg 2000a, 2002 and Ries 2001). The structure of the basis functional network can be found immediately by simple experiments. One has only to observe the direct succession of words in large text collections: It is obvious that in text usually each word (form) is repeated several times and is located at various places. But in a network there exists only one node for each such word. If in text for every word all the different direct successors and predecessors are determined – we used large text collections of linguistics, so-called corpora with millions of running words – then at the same time the connections of the corresponding nodes in the network to any possible successor nodes and predecessor nodes are known. Thus the complete structure of the network at the basis level is immediately derived. Large numbers of nodes and much larger numbers of connections result. For large text collections typical amounts of vocabulary are several hundred thousand different words (to be precise: word forms). For estimations there is an interesting result: It could be found that measurements with *N* different words in a given text had about 5*N* different direct word successions, which are equal to the same number of possible word pairs (Meyer 1989). In an experiment, using LIMAS-Corpus, one of the most familiar German text collections, about *N* = 120 000 different words can be found, which gives the same number of nodes in the basis network of the model. It is not possible to represent this large network in the form of well-known mathematical graphs. Another representation is necessary.

A suitable method will be explained first by a small example. In Fig. 2a one of many possible examples for a graph is shown. It is the Siedenburg-graph (Siedenburg 1992) with its nodes and connections (We like this demonstration object with no direct reference to language problems, because it is a simple mathematical model. Nevertheless it has more favourable properties than the well-known hypercube. The graph was invented in our research group as a model for superior computer networks several years ago.). Below this graph one can find in Fig. 2b a more abstract representation, which is called "connection matrix". Here nodes were given numbers and these are placed along both axes. If there is a connection between two nodes, then only a dot is entered in the matrix. The number of intersecting connections in a network is thus reduced to a corresponding number of simple dots in the matrix, which gives a clear result, especially for very large networks. It should be noted that the complete number of dots in the matrix describes exactly the complete network structure.

In the case of text measurements, following the principle of Fig. 2, a structure is obtained as is shown in Fig. 3a,b. Here the words were listed in the order of their frequency of occurrence, where the position in the list is called "rank r" and plotted at the axes. (A more fundamental and structural definition of rank in networks is given later on. It has the same number scale). The peculiarity in Fig. 3b is the logarithmic scale of the axes, deviating from the simple linear scales in Figs. 2b and 3a. This scaling and its use in linguistics was introduced for the first time by G. K. Zipf for statistical purposes in 1949 (Zipf 1949; Li 2002; Guiter, Arapov 1982). Since that time, especially in quantitative and statistical linguistics, this has become the usual method of representing the results of measurements. By the logarithmic scaling we can read the growing amount of word numbers (or rank numbers) in equally long sections on the axes, with a compression given by a constant factor, e.g. by a factor of 2.

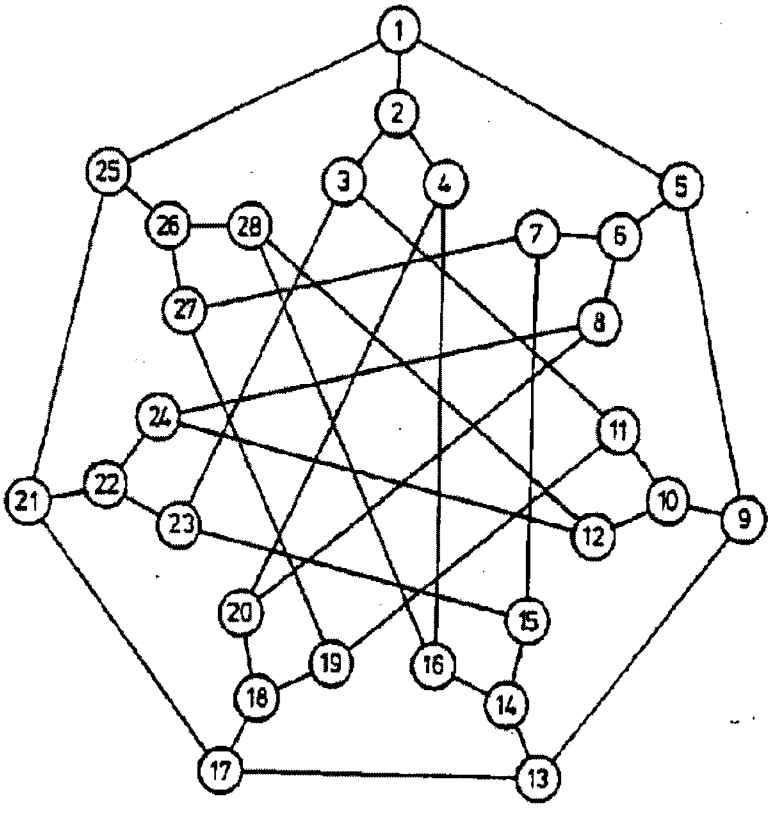
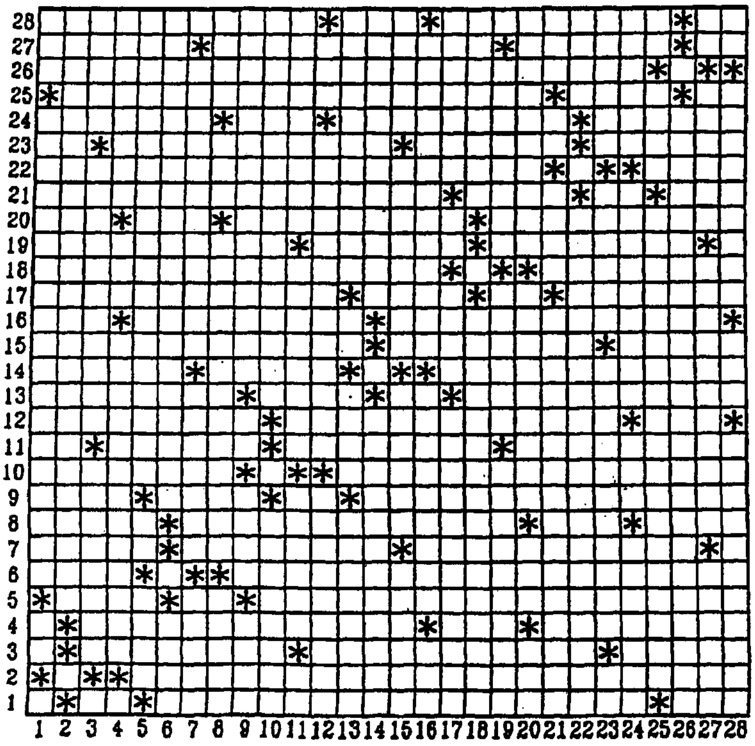


Fig. 2a. The Siedenburg-graph, with 28 nodes and 3 connections per node, depicted in the usual way



28												*				*										*		
27							*												*							*		
26																									*		*	*
25	*																				*					*		
24								*				*										*						
23			*												*							*						
22																					*		*	*				
21																	*					*			*			
20				*				*										*										
19											*							*									*	
18																	*		*	*								
17													*					*			*							
16				*										*														*
15														*									*					
14							*						*		*	*												
13									*					*			*											
12										*														*				*
11			*							*										*								
10									*		*	*																
9					*					*			*															
8						*														*				*				
7						*									*												*	
6					*		*	*																				
5	*					*			*																			
4		*														*				*								
3		*									*												*					
2	*		*	*																								
1		*			*																				*			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28

Fig. 2b. The same graph shown as a connection matrix

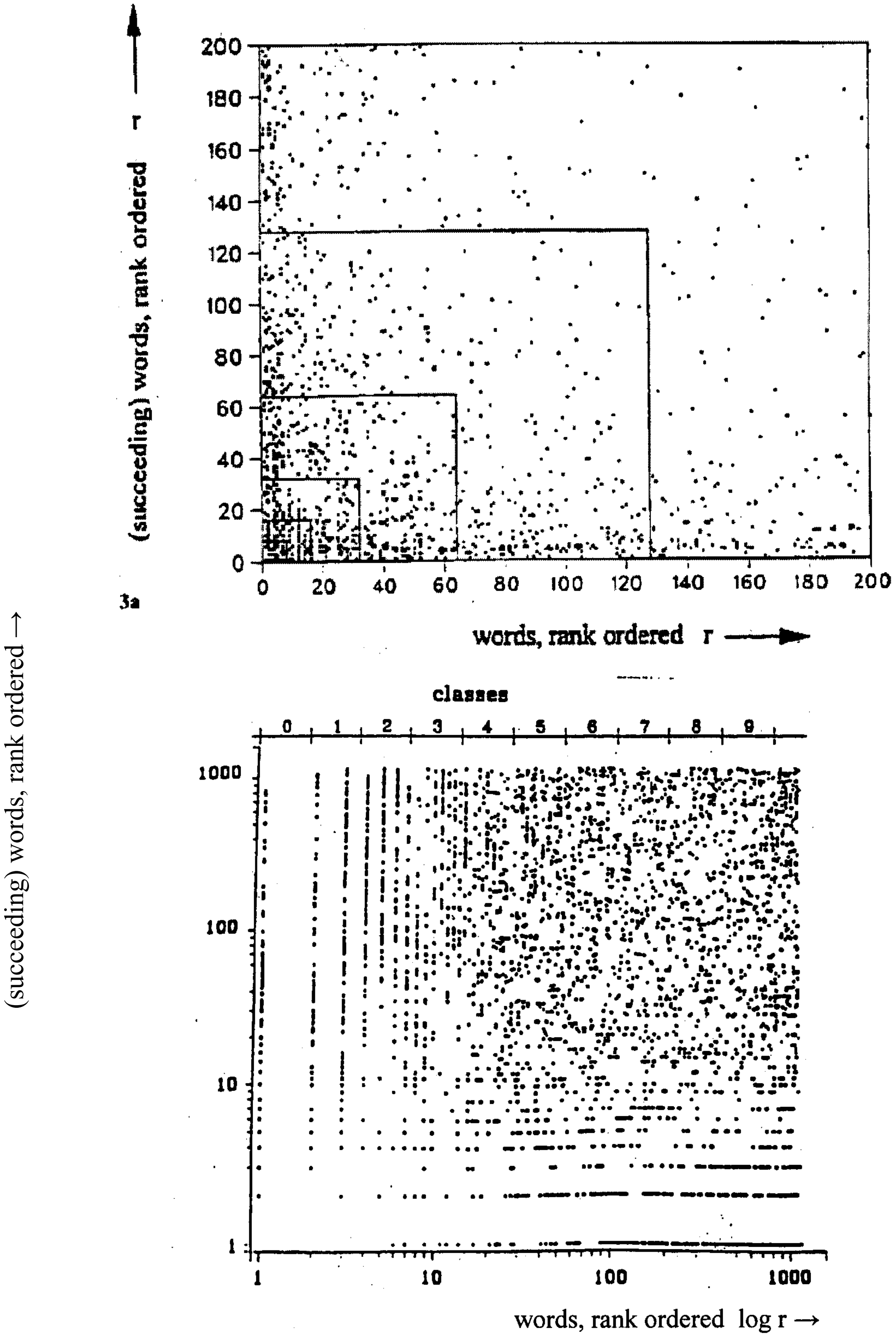


Fig. 3. The connection matrix representation of a graph which was derived from given German text. (It is not a random distribution of dots.) Fig. 3a shows the matrix, where words are rank ordered with linear scaling. Fig. 3b shows the same object in a matrix with logarithmic scaling. Grouping of nodes in classes is indicated at the upper side. In such a matrix, a nearly constant dot density can be found in good approximation.

Comparing Fig. 3a with Fig. 3b it can be clearly seen that at the axes the rank numbers and therefore also in the matrix area the corresponding dots are accumulated more and more. (In Fig. 6a the logarithmic compression of rank numbers will be illustrated again in another way).

5. The association matrix

The new diagram in Fig. 3b was not yet known to G.K. Zipf because it does not describe statistical facts. Instead it shows invariant connections of nodes (here they are still ordered by a statistical parameter that will be replaced by a structural parameter later on) and was published first in 1988 under the name "Association Matrix" (Hilberg 1988). The matrix should be interpreted in the following way: When seizing a first node with its number upon the horizontal axis you will find dots on the vertical line above this numbered node. Proceeding from these dots horizontally to the left, you will find on the vertical axis the numbers of all nodes which are proven possible successors to the first node at the beginning. On the contrary, if one is looking for predecessors for an arbitrarily chosen word, one has only to go to its number on the vertical axis and then you will find on the horizontal line dots which lead downwards to all possible predecessors. Of fundamental importance is the order of node numbers which defines their rank. Zipf decided that at the x-axis the order is chosen according to the frequency of occurrence of words. Following provisionally this proposal also for the matrix in Fig. 3, the most frequent word is placed on the horizontal axis at the farthest left end, and going to the right, less frequent words follow, until at the farthest right side those words can be found which occur only once. (In Zipf-diagrams, contrary to the association matrices above, not rank but absolute frequencies are plotted in the vertical direction).

Investigating structures of natural language text we were surprised to find an unequivocal relation between frequency statistics and deterministic network structure. That is, we found that rank is an order which is identical to another order given by the number of connections which a node can have with other nodes (which we called "ramification"). This fact is by no means self-evident. The remaining difference between rank and ramification is only a reverse direction in the axes when counting numbers. That is, the word occurring with the highest frequency (it has lowest rank) is the same as the word in the node with the greatest ramification.

6. Dot distributions and various texts

Now, having obtained experimentally a first reliable image of the functional structure of the human brain in the level of words, the question arises how we can understand and interpret the language network of Fig. 3b. First of all the nearly constant dot density over the matrix area is very striking. It looks as if somebody had dispersed grains of sand over this area. Yet we know that each dot does not represent a random succession of two words in text, because a dot stands for an existing and proved network connection (we deal here with classical deterministic linguistic conceptions and not – as very often in contemporary science – with probabilistic or statistic conceptions). Moreover the matrix is sparsely occupied because it is obvious that there exists a much larger number of theoretical successions of words, which are not allowed in language and which never will get a dot.

If we derive a network from another German text Corpus or from works of various German writers, e.g. of Martin Luther, Friedrich Nietzsche, Johann Wolfgang Goethe, Thomas Mann, etc., or from various Corpora of newspapers, e.g. of FAZ, Spiegel, Süddeutsche Zeitung, etc., then exactly the same distribution of dots results, provided that we used a sufficiently large vocabulary. Word connections, permitted by the language community (only direct word successions) will always be found at the same place. When proceeding from Corpus to Corpus, some new word connections are added, but they will become increasingly seldom, according to experience. The corresponding additional dots for these connections are placed at rows and columns between previous dots. Only when the vocabulary is extended by new words, additional rows and columns will arise, which will slightly distort the image of dot distribution locally. At last also these perturbations can be avoided by using a modified chronological rank.

7. Fixed dots and word classes

A completely different individual distribution of dots results when we use text of another language. Any language has a network structure of its own, a fact that can be demonstrated very easily. At the same time it is true that for each language, provided we have very large text collections, the association matrix is invariant. It is astonishing how different languages are equal in their constant dot density and at the same time how they differ in the fine local placement of dots. The conclusion can be made that this is true for all languages which exhibit a regular Zipf-curve. This again is proved practically for all natural languages. In all the languages investigated by us, the distribution of dots inside the matrix seems to be statistical or stochastical at first sight, possibly causing the general misunderstanding that this would obviously represent a random result – a conclusion that may be drawn by inexperienced observers or by routine blinded scientists, perhaps coming from statistical language processing (Manning, Schütze 2001).

A comparison of a large matrix full of dots with the sky full of stars may illustrate the proportions furthermore: The position of stars could at first sight appear random and even variable to a naive observer, but when longer observation demonstrates to him that stars do not alter their mutual positions, he will surely understand the name "fixed stars". In this respect one could also speak of "fixed dots" in the matrix.

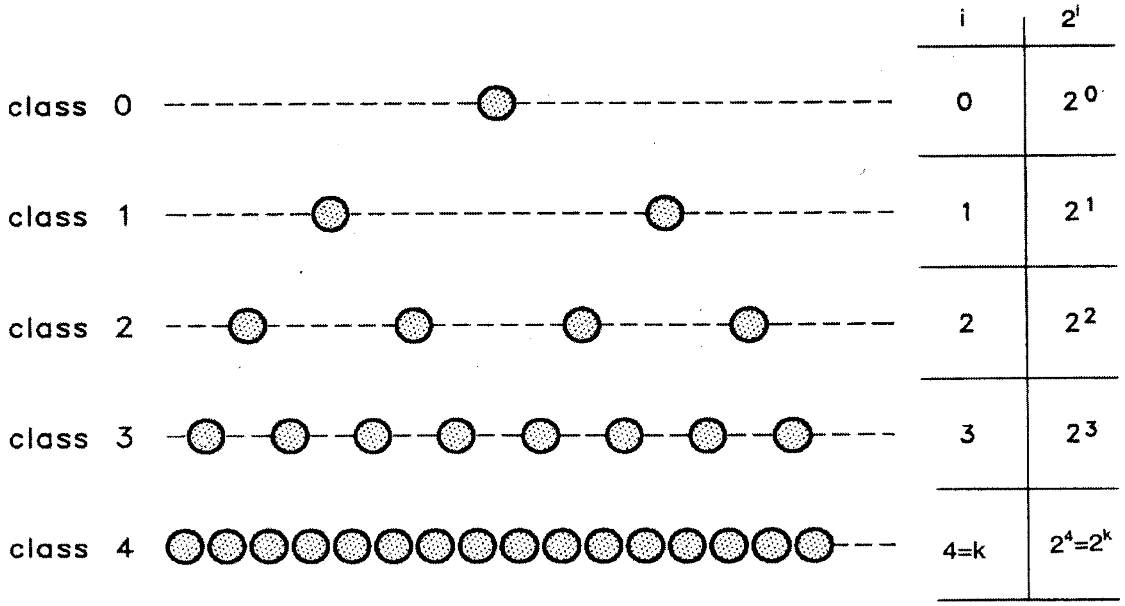


Fig. 4. Illustration of a basic model where nodes are divided into word classes. There are classes i = 0,1,2,…,k, each exhibiting 2^i nodes. Each class has nodes of equal ramification. The number of nodes in a class multiplied by the number of ramifications equals the constant value 2^n.

We have learned now that the fundamental structure of functional networks is the same in all languages. A more precise understanding of common properties of all language networks can be obtained when the matrix area is divided into narrow horizontal and vertical stripes (see Fig. 6a). Considering constant dot density and logarithmic scaling, a structural classification of words or nodes, respectively, is adequate, see Fig. 4. Here, in a first approximation, the number of nodes in a class – beginning with one node – increases from class to class by a factor of 2. At the same time the numbers of connections per node decrease correspondingly by a factor of 1/2. The product of both numbers in a class is a measure for the communication performance of the system under consideration. It is the same for all classes (it corresponds with the fact that in the matrix the number of dots in all stripes is about the same).

8. Shannon and Zipf

In the field of linguistics the following will be of interest: For the last five decades in the preceding century the well-known law of G.K. Zipf (1949) has been in existence. This brings, in a mysterious way, a clear mathematical relation into the apparently free world of man-made text which is proven to be valid for very differing types of text samples and for very different languages. The description of the law is very simple: As was mentioned before, when arranging frequencies of words in a diagram with logarithmic-scaled axes, one for frequency and one for rank, Zipf's law can be recognized as a straight line declining by 45 degrees. A lot of these diagrams have been measured by many researchers in the past and also by ourselves (Hilberg 1999, 2000b; Meyer 1989; Steinmann 1996; Nachtwey 1995; Burschel 1998), because apparently nobody can trust this mysterious law until he himself has examined it. Instead of presenting our own recent results, Fig. 5 shows the most popular example in natural science literature given by C.E. Shannon, the well-known founder of information theory. Such a clear penetration of mathematics into the amazingly flexible world of language, especially in litera-

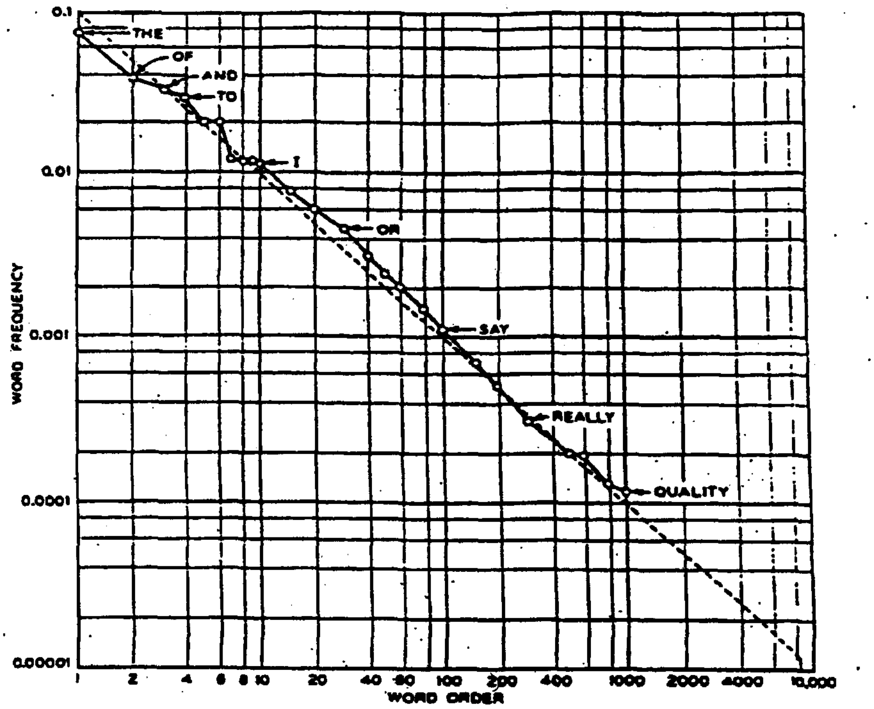


Fig. 5. Example of a Zipf-curve, which was already published by the mathematician C.E. Shannon (1951) 50 years ago. Such a curve with slope -1 can be found in any text of any language.

ture, appeared unexpected and impossible to all researchers, see Table 1 in Appendix 4. The underlying cause could not be understood – even Shannon had no explanation but nevertheless he used the law! However, by means of the results of our matrix measurements above, we shall approach the mystery in a first step in the following way: We supplement our matrix measurements by including also the frequency of dots, i.e. how often a certain word pair (dot) appears in text, and we store this data for all dots, e.g. in a third coordinate. Then these frequency numbers are summed up for each word in its matrix column. As can be seen, this method combines structural determinism (the individual connections and the resulting ramification of nodes in a network used as rank order) and conventional statistics (frequency of words in text) and yields in the average exactly the famous Zipf-curve. As was also mentioned earlier, the Zipf-relation is the base of classic quantitative linguistics, although nobody understood at the bottom what could be the reason for such a power law which is present in any text in any natural language; compare for example the thoughtful considerations of the American Nobel prize winner in physics, M. Gell-Mann (1994) ("...Zipf's law remains essentially unexplained..."). However, this situation is going to be changed. We shall reach an understanding of the necessary network structure and of the root of Zipf's law with the aid of the class model in Fig. 4 by further steps and evaluations which are to be explained subsequently in detail.

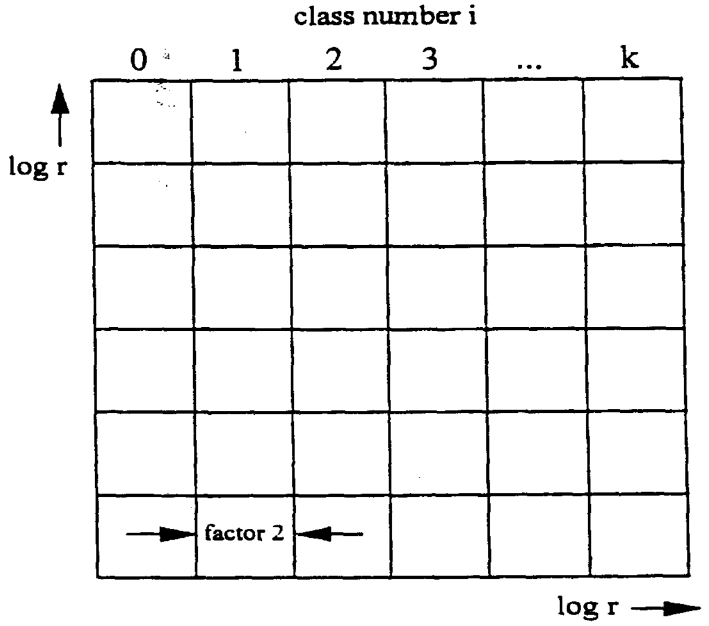
9. Formal design of the superior general type for the language network

In the preceding chapters the object was measuring and describing language networks of a natural language (with the restriction to direct successions of words). The structure of the network discovered was not known in mathematics up to now, which seems very strange. Furthermore it is another interesting question, whether the technical design of such a network with its connections is possible and whether it can be realized at all, so that we may be able, for example, to construct a corresponding electronic language machine. Now, if for once we could temporarily forget the deterministic character of the network structure in the matrix, and if we would interpret the seemingly random distribution of dots as a true mathematical distribution by chance, it could bring us very interesting information theoretical insights. That is, it will lead us at first to a well known general law of information theory which is valid for all random types of such networks, a law however which has to be observed in fact also in comparable optimum pseudo-random deterministic networks. Therefore the tentative interpretation as a random distribution of dots in the matrix will indeed receive a subsequent justification.

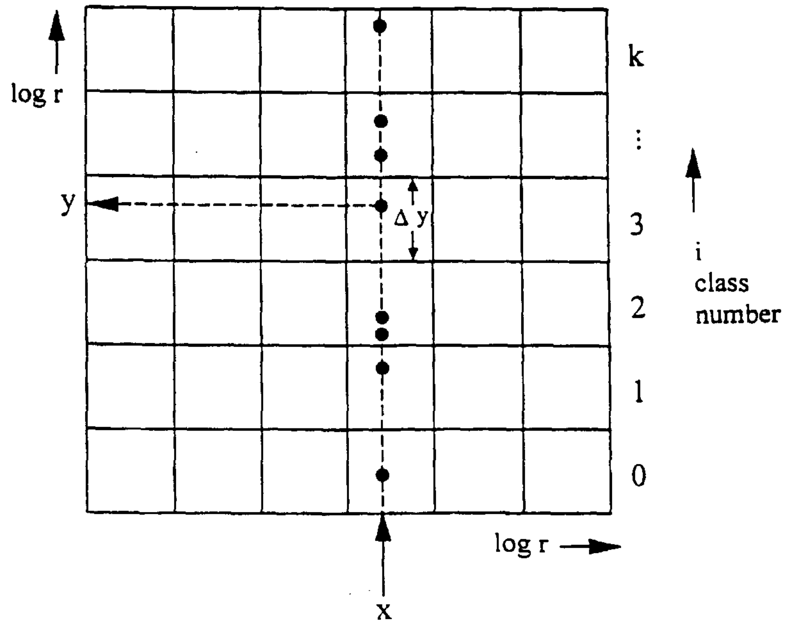
10. Random distribution of dots

Now we need a suitable method for generating a random distribution of dots with constant dot density in the matrix area and with logarithmic scales in the axes. That is not very difficult, as can be explained in Fig. 6. In the simplest case we have only to presume a class division of nodes as in Fig. 4, corresponding to a given large amount of words (vocabulary) with rank limit $r_{\max} = N = 2^{n+1}–1$. At first we assume that these nodes do not have mutual connections, we know only their rank numbers. That is, contents and meaning are not defined at the beginning. The following procedure generates a network which is characterized by a random distribution of dots in the matrix and especially by a type of dot distribution of constant density which can be observed also for natural language networks.

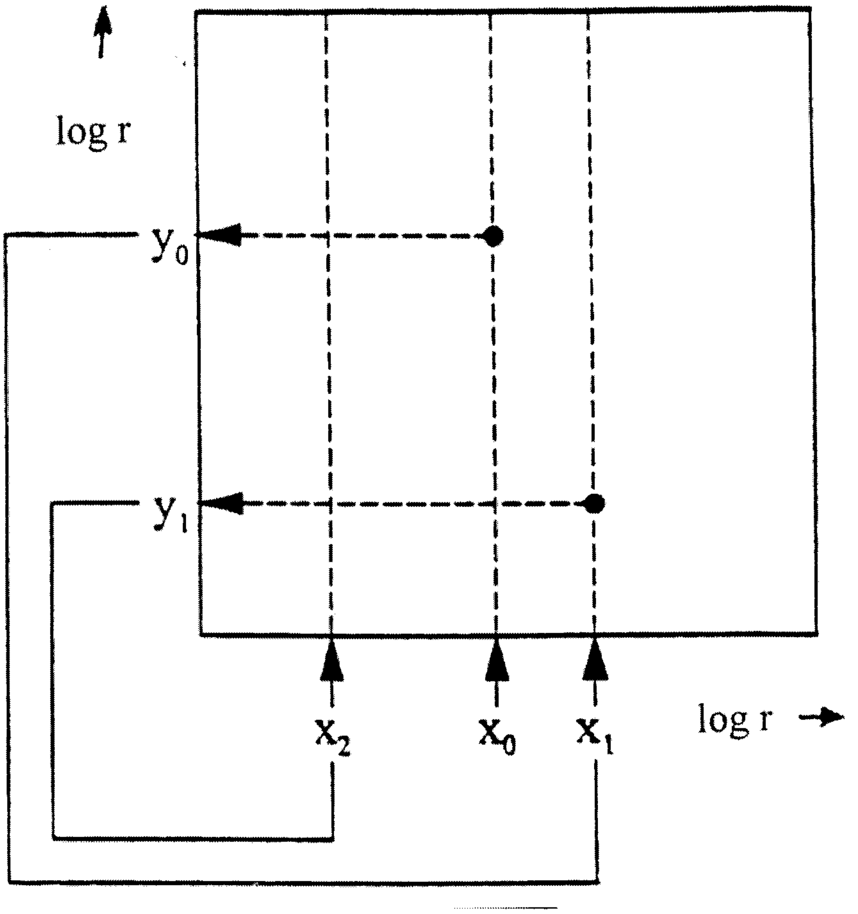
Fig. 6a reminds us at first what the logarithmic placement of nodes with rank numbers at the axes means and how the classification of nodes will divide the matrix area in narrow stripes. Then the design begins in Fig. 6c with an arbitrary word node x_0 (at the lower edge of the matrix), for which the connection to the next word node y_0 (at the left edge) is to be determined. This word node y_0 is situated of course in one of the structural word classes. We decide on one of them by chance, that is, we choose in Fig. 6b one out of k+1 available horizontal stripes by chance (in mathematical terms all classes have the same probability). Classes i, with i = 0,1,2,...,k, contain an amount of word nodes, the number of which increases with 2^i. Inside a chosen class we choose again one of its words by chance. This is the desired word y_0.



6a)



6b)



6c)

Fig. 6: Sketches for the explanation of a construction process in two steps which yields a network of randomness.

6a: It is shown how the matrix can be divided into small vertical and horizontal stripes with width factor 2. The number of different words at the edge of the stripes is the same as the number of word nodes in corresponding word classes (indicated by points at the edges)

6b: Principle of random choices of a class and subsequently of choosing a dot in this class, starting with word x and ending with the next word y.

6c: It is shown how an iterative process proceedings occur from nodes x_0, x_1, x_2, ... to the successor nodes y_0, y_1, y_2, ... with $y_0 = x_1$, $y_1 = x_2$, ... In this way the matrix area is filled with dots.

In the next step we could link the nodes x_0 and y_0 in a graph by a connecting line (not shown here). The following second connection line will be found by repeating the whole action in the matrix. That is, now beginning in Fig. 6c with the last point y_0 – which as a new starting point obtains the name x_1 – we choose again one of the classes by chance and inside the chosen class one of the possible word nodes by chance. Then we obtain y_1. Repeating this action sufficiently often, e.g. until a sufficiently large part of all word nodes has been called up at least once, a random distribution of dots results as is shown in Fig. 7. It represents a randomly-generated individual network. When after its generation no alteration takes place, we can call it a firm network that in its general type corresponds with the biological functional language network in the brain of an adult human. Statistically seen the artificial network cannot be distinguished from a natural language functional network (e.g. it has again apparently statistical symmetry of the dots to the rising diagonal and it exhibits maximum entropy, which will be explained later on).

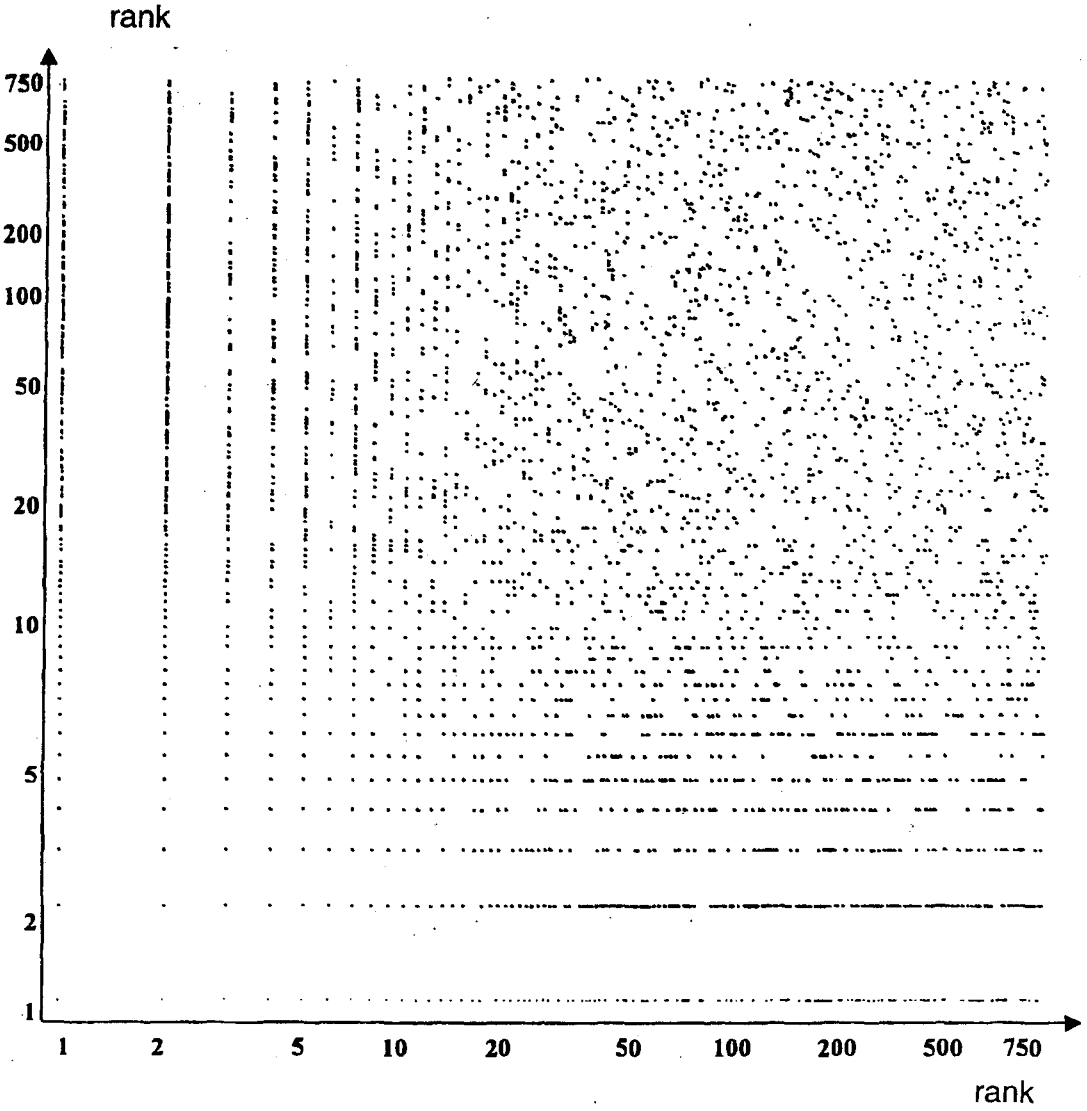


Fig. 7: The result of such a stochastic experiment. It describes a network of maximum entropy

How fruitful this result of a formal network design is may be recognized by a statistical evaluation of the model in two tests:

Test 1: Observing natural text it can be seen that the frequency of punctuation marks inside this text is rather high, especially that of commas and full stops. Summarizing full stops, question marks and exclamation marks between sentences in a single symbol "separation mark" and neglecting commas for the present, we shall find this general separation mark in the first word class, which, as we know, can theoretically have only one element. We utilize such a class in the design of a random network. Then the following fact is valid: When all classes are activated with equal frequency, the class of general "separation mark" will appear from time to time. A sentence is always terminated with its appearance. Having k+1 classes, $n = 0,1,2,...k$, the separation marks appear with probability 1/(k+1). Between the marks words are present. A sentence includes therefore k words on the average. The encyclopaedias of German language cite that the most frequent lengths of sentences amount to 15-18 words, with a high value of about 16 words in the newspaper "Frankfurter Allgemeine Zeitung" and a low value of about 11 words in the newspaper "Bild". Regarding voluminous collections of text, e.g. LIMAS-Corpus with about 120 000 words and $2^{17} = 2^{k+1}$, then $k = 16$ classes can be counted. Substracting the class for separation marks these numbers correspond to an average word length of 15 for a sentence. Considering that we deal only with orders of magnitude, the precision of the estimation taken from the association matrix is very surprising. Also surprising is the fact that we derived this result without the lengthy statistical measurements of long text that was always necessary in the past (Meyer 1989).

However the model gives no explanation, why above the volume of a usual supply of words the average length of sentences continues to grow with the number of different words, though for very large vocabularies the growth becomes extremely weak because of the logarithmic scale. Maybe there is a connection between very rare words or segments and long explanations, or there is an influence coming from authors like Thomas Mann or from increasingly complex scientific literature.

Test 2: If, in the random design process described above, we accumulate the dots in Fig. 7 in the columns, a steadily declining curve is obtained, which is called "ramification curve", because it gives the resulting ramification number for any node. This straight curve has similarity with Zipf's curve. However, only when we count the frequency of the activated nodes additionally during the generating process and store it in a third coordinate (as we have done above for natural text) and when we accumulate the frequencies of the single dots in the columns, now as before in dependence of ramification rank in a diagram, eventually a curve is obtained which is at the bottom identical to a measured original Zipf-curve, see Fig. 8. Following the rules of its design this curve descends precisely like a staircase. Simulations showed that it is somewhat smoothed by Gaussean-like distribution effects (Steinmann 1996). The mean slope of this curve has necessarily the value of -1 , that is a descending slope of 45 degrees. The logical reason for this is obvious: All classes are chosen with equal frequency, whereby the number of nodes in any class i rises with 2^{i}. Therefore the probability for an arbitrary single node in a certain class i is only $1/2^{i}$. (Second order effects and further improvements of the model were carefully calculated by F.-M. Steinmann (1996)).

As is well known, the conventional approximation of the measured staircase curve is usually shown in a first approximation as a straight line with slope -1 or in a second approximation (Guiter, Arapov 1982; Rapoport 1982) as a curved smoothed line (curve fitting). However the model curve with correct stairs is far more precise. Strictly speaking, Zipf's curve is not a pure power law but a staircase law in reality. A model that produces this behavior should be superior.

11. The unexpected mathematical core is maximum entropy

Probably the statement is correct that any natural language has a firm and individual network of its own. By means of random design processes we can also generate many networks of the same type and, if we like, we can begin to create new language systems. It may be possible, though not very probable, when we repeat the design process very often, to find eventually structures of known natural language networks. More realistic seems to be the question whether natural networks were also once created in this random manner. Thus, who created languages and was he perhaps really playing dice? Or, to ask within the scope of natural science: Why are laws of randomness or laws of pseudo-randomness built into these networks? At first we understand that the elementary structure of classes in Fig. 4 has a simple physiological basis which can frequently be found in nature. This basis consists of cell division processes where the number of new cells multiplies always by a factor of 2 . Now the only problem that is left to us is eventually to clear up the crucial process of realizing connections between network nodes.

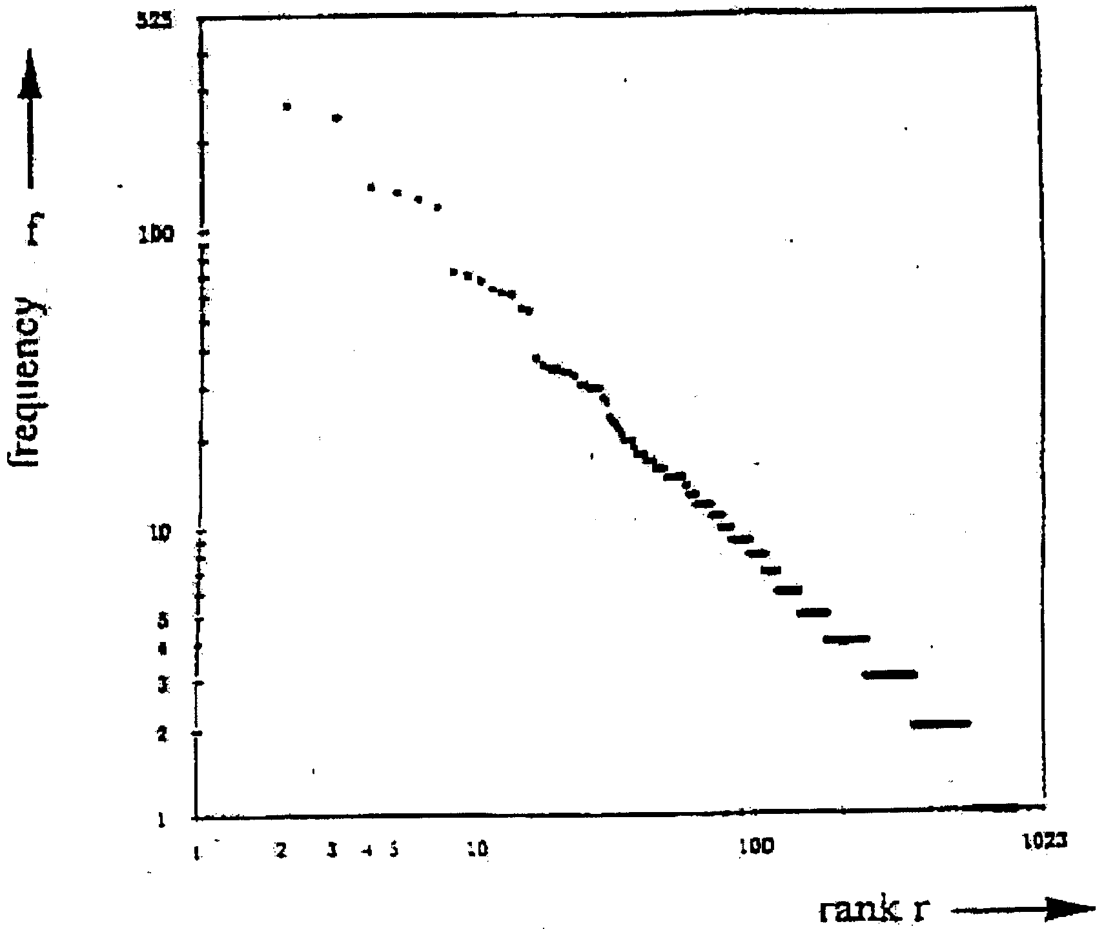


Fig. 8. A Zipf-diagram which can be derived from the stochastic experiment described (see Figs. 6 and 7). The more nodes are used, the better is the slope approximated by the value -1.

In recent years scientists have learned that the connections between brain cells or neurons (with their axons, dendrites, synapses, etc.) develop over a period of only a few years in little children. The connection patterns depend on the language community in which the child grows up, or in other words, the contemporary language defines the structure of connections, see for example recent affirmative research results (Dehaine-Lambertz 2002). A pseudo-random law is built into language, as was shown. If language is an invention of man, how could a random law arise? We should learn about this from the development of artificial probability networks above. Here we made use of the simplest principle of all, the principle of randomness, in order to find the connections between nodes. Thus no human geniuses were necessary. In spite of this, however, even geniuses could not have acted more cleverly than to choose the principle of randomness. This can be understood when the entropy of text is calculated (entropy is the average information content of symbols; words are here regarded as symbols). With this operation the frequency of words in natural text is conventionally counted and introduced into the well-known formula of entropy, originated by C.E. Shannon. Equally satisfactory is the calculation of entropy by utilizing all text paths in the language network, which were laid down when the network was created. Both values of entropy coincide remarkably well, though one of these calculations – the conventional one – is based on a lot of measurements of word frequencies in given text, and the other calculation – the connectionistic one – requires only the number N of involved nodes in the network (Steinmann 1996). The most important discovery is, however, that by constructing a network by chance (following the principle of Fig. 6, which leads in the end to a nearly constant dot density) eventually the maximum of all possible entropy values is exactly obtained (Hilberg 2000b). (As is well known in information theory, the maximum of entropy results, when words – being here the information symbols of the signal source – can be predicted in any output action only with maximum uncertainty. That is exactly the case when throwing dice. In most applications the dice is thrown only once for each output action, corresponding to a one stage signal source. Here however in a two stage process for every decision the dice is even thrown twice. Not directly worse than doing it only once. It is rather a generalisation for multistage sources (Hilberg 2000b).) This means in fact that networks with the structure described above are able to deliver word sequences of maximum information content on the average. Conscious human creators of a natural language could not do better.

Finally this has shown the superiority of the pseudo-random structure of networks over all other possible structures. In the view of mathematics, maximum entropy is the unexpected deep lying core of natural languages. Its implication is the mathematically defined network structure. One could imagine that in the evolution of languages the pseudo-random structure was not present from the onset but that it was slowly improving thanks to the endeavour of many humans. This may have led gradually to an optimum of efficiency. The combination of the simplest instruction rules for generating language and the achievement of optimum information properties may be the explanation for the striking success of "homo sapiens" compared with other less talented competitors.

12. Further investigations

We observed above a language-network only in the abstraction level of words. But, of course, human memory can store far more information than only a lot of single words and the possibilities of their direct succession. In order to understand language we still necessarily need the so-called "context", which is included in sentences and the succession of many coherent sentences. If we interpret the immediately observable language-network of words merely as the surface structure of a complete system (on this surface only the words can be grasped directly, i.e., they can be heard, spoken, written, read, etc., something which is no longer valid for thoughts), then a deep hidden structure is still missing. This cannot be determined by the kind of measurements described above, it can probably only be invented by clever ideas. One possibility is seen, following the principle of increasing abstraction in steps (Hilberg 1997a, 2002a), in creating a model containing several language networks of the optimum type discussed above (Hilberg 1997a, 2000a). This is indeed a new approach, because in the scientific mainstream dealing with conventional artificial neural networks the existence of a hidden deep structure is extremely dubious, as can be cited from Thompson (1994): "perhaps in brain, deep language structure will not exist."

In order to build an intelligent language machine exclusively out of functional networks, a structural machine was designed which is strictly contrary to the von-Neumann-computer model with its combination of large different modules. It consists of a hierarchical arrangement of network layers of increasing abstraction. (The method of systematic consideration of extended context in text segments of increasing lengths can only be sketched, for details see, for example, Hilberg 1997a, 2000a, 2002.) The basic idea again is very simple. If we intend to proceed along a certain text path in a given language-network, which will generate the output of a corresponding text as a series of words, see Fig. 1, the necessity arises that the transition from one node to the desired next node must be chosen using a suitable control signal. Such control signals, called meta-words, will come from a so-called *meta-language-network* where the signals are stored in compressed form. (In order to avoid an exploding system the requirement was laid down by us that the number of meta nodes must not be larger than the number of words in the basis word level (Hilberg 1997). This can be met without any restrictions by an efficient abstraction process during the time when text is received and in the reverse direction by a logical prediction process, when text is generated. For comparison, in the field of statistics, prediction would be performed by well-known Markov processes but here the procedure works without probabilities.) The paths in this meta-language-network will be chosen again by control signals of the next network, *the meta-meta-network*. In this way the process will be continued over several hierarchical abstraction levels until we arrive at the last level with its relatively short meta-word for the whole original path. By means of this high level meta-word we are able to recall the whole long text again and bring it to the output, where we can read and hear it in the familiar word forms. (An example for such a text, generated from a higher situated single meta-word can be given here as a small section in German (Ries 2001): "...unterm breiten Dache sprudelt ein prächtiger Brunnen, spielte den blanken Fenstern stehn einige Blumenstöcke...". There is still a small error in this text because the necessary deterministic prediction procedure used here was apparently chosen too weak.) Logically, the last meta-word is the redundancy-free representation of the supplied text. We like to speak of the "thought" of this text.

The Darmstadt research project runs under the label "language machine". We understand this text-machine in some respects as a counterpart to the computer, for it should handle words better than numbers. Moreover it should take up – very fast and with little technical effort – reasonable text of given length, which has to be stored and, if desired, again to be handed out exactly word for word or in modified form. Future applications may be seen in the field of ambitious translation of text from one language into another language. Present preliminary experiments have shown that a simulated system was already able to deliver text, which had first been received, in various modifications without using any grammar rule (Ries 2001). In future the spectrum should be enlarged so that a choice can be made from a relatively unsharp report, as one would expect from an untrained person, to an exact word for word repetition as it is given e.g. by a digital memory.

We think that a lot of work is still necessary until we will achieve a mature state. A first survey of the principles of such an intelligent language system can be found in Hilberg 2000a, special aspects in the papers Hilberg 1997a,b, 1988, 1990, 1999, 2000b. Work has proceeded so far that several levels of abstraction were realized and that execution and cooperation of particular connectionistic operations for text generation and text processing could be verified in simulations, such as abstraction, compression, prediction, segmenting, etc., see the doctoral theses Meyer (1989), Steinmann (1996), Nachtwey (1995), Burschel (1998), Ries (2001).

Acknowledgement

I would like to express my thanks to Prof. Dr. Peter Rechenberg, Computer Science Department, University of Linz in Austria, because his many critical questions led me to make corresponding improvements, which, I hope, finally resulted in an article which can be understood.

Furthermore my daughter Eva fortunately returned just in time from an Erasmus exchange year in England so that she could help her father with the translation of complicated text passages – to my regret the translation machine was not yet working!

Appendix 1

The question why the results of randomly generated text are equal to the results of natural text is not yet answered rigorously. We may approach this answer again in some steps. The remarks made to Test 2 above considering the origin of Zipf-like curves in artificial language networks can be applied also to a measured natural network. As a first step one can imagine that a text path is chosen by chance in a given natural network, just as we did above, when we played dice without boundary conditions. It is as if we are now walking aimlessly and randomly in a "maze" on given rails (where for example the German text arises of the following kind: " ...so solches nicht mit Fluchen und gleich auf ihn der nach Arbeit sie mahnte...").

If once again we store in every matrix dot the frequency with which we went through each corresponding connection, the addition of all frequency numbers in each matrix column will lead to the frequency of this word, which was encountered in text, independent of the fact which special ramification after this node then took place. This frequency is the same as that of words in a Zipf-diagram. Therefore we again have a Zipf-curve and it has necessarily a mean slope with value -1 .

Last but not least we have to investigate the true classical Zipf-curve (be cautious: the conventional interpolation with formulae of increasing precision which can be found in literature has nothing to do with logical derivation, it is only curve-fitting). The difference between generating reasonable text and the playing dice considerations above is that in the brain we do not walk by random decisions in a natural network. The text path is rather chosen by regarding context information which is present in coherent text. This is the usual situation when measuring original Zipf curves. In this case we have to run through the functional network with a reasonable text and in so doing, we shall meet the same nodes as encountered with random choices but in another order of succession. The same frequency values will result. (Remember also the lottery game in the beginning). Again: the main difference between a coherent text and random text is the fact that in the first case relations exist between words, also at a remote distance. These relations, however, are not considered in a Zipf analysis, where only the frequency of occurrence of words is counted. Thus without further knowledge, the movement along a reasonable text path cannot be distinguished from walking by chance through the network along existing network connections. Far reaching relations which are necessary for a reasonable text can be considered only when appropriate structures in additional networks above a basis network are introduced. The adjacent network levels are called "meta levels" and they are organized in a hierarchical manner.

Appendix 2

The linguistic problem in test 2, where we were interested in discovering the relation between network structure and Zipf's law, let us enter the field of statistics. As is well known, statistics has to do with counting events, and the measuring of the frequency of numbers is a typical statistical operation. In contrast: The association matrix with its pseudo-random distribution of dots (without additional frequency values) has the meaning of being a firm network structure and therefore it is not an object of statistics! It is rather the basis for future connectionistic artificial language systems.

Appendix 3

A supplement seems to be necessary in order to separate notions used here from notions used elsewhere. The considerations above have shown that the new technical connectionism prefers higher abstraction levels with physical properties, in contrast to the classical concept of low level layers of detailed artificial neural networks. The new connectionism is characterized even on the basis level by the use of stored codes in the nodes of a model network. The structure of this model is derived from measurements – that means it is learned from nature! Furthermore, the connections between these nodes can be switched on by a single individual learning operation and after that the connection will essentially persist forever. Thus a network arises which is deterministic and well known in its local details, in contrast to the classical neural network which has to be treated as a black box, in which all weight values at the input of all neurons are adapted in a statistical way to meet given requirements. The well-known specific back-propagation-algorithm shows that the system behavior defines magnitude and distribution of all weights. Consequently, in the course of a learning process or when the task has been somewhat changed, the previous weights may be considerably and simultaneously altered at many places, with values either increasing or decreasing. This is definitely not the case with the networks in higher abstraction levels, which were considered here.

Appendix 4

Table 1
Researchers who were engaged in solving the enigma of Zipf΄s law

G.K. Zipf, 1949 **Linguist**
(Human behavior and the Principle of Least Effort)
C-E.Shannon, 1951 **Mathematician (Information Theory)**
Prediction and Entropy in Printed English
B. Mandelbrot, 1953 **Mathematician (Chaos-Theory)**
An information theory of the statistical structure of language
A. Rapoport, 1982 **Mathematician (Philosopher)**
Zipf΄s Law Re-Visited
W. Li, 1992 **Information Scientist**
Random Texts Exhibit Zipf΄s-Law-Like Word Frequency Distribution
M. Gell-Mann, 1994 **Physicist (Nobel-Prize)**
The Quark and the Jaguar. Adventures in the Simple and the Complex.

References

Bassenge, G. (2001). *Automatische Klassifizierung von Wortformen in Texten der deutschen Gegenwartssprache*. Diss. TUD.

Burschel, H.-D. (1998). *Die meßtechnische Ermittlung von Assoziationen zwischen Worten in kohärentem Text und ihre Nutzung bei Prädiktionen verschiedener Reichweite. Diss. TUD.*

Dehaine-Lambertz, G. (2002). Babyphone. *Science 298, 2013* (Süddeutsche Zeitung 10. Dec. 02, p. V2/7).

Gell-Mann, M. (1994). *The Quark and the Jaguar. Adventures in the Simple and the Complex*. New York: Freeman.

Guiter, H., Arapov, M.V. (eds.) (1982). Studies on Zipf's Law. (= Quantitative Linguistics, Vol. 16). Bochum: Studienverlag Brockmeyer.

Hilberg, W. (1988). Das Netzwerk der menschlichen Sprache und Grundzüge einer entsprechend gebauten Sprachmaschine. *ntz Archiv Bd. 10, H.6, 133-146.*

Hilberg, W. (1990). Der bekannte Grenzwert der redundanzfreien Information in Texten – eine Fehlinterpretation der Shannonschen Experimente? *Frequenz 44, 9-10, 243-248.*

Hilberg, W. (1997a). Neural networks in higher levels of abstraction. *Biological Cybernetics 76, 23-40.*

Hilberg, W. (1997b). Theorie der hierarchischen Textkomprimierung. Informationstheoretische Analyse einer deterministischen Sprachmaschine.Teil I. *Frequenz 51, 7-8, 196-202;* Teil II: *Frequenz 51,11-12, 280-285.*

Hilberg, W. (1999). Mandelbrot´s Gesetz der maximalen Entropie in natürlichen Sprachen als Folge der Struktur des neuronalen Sprachnetzwerkes. *Proceedings of the Workshop on Physics and Computer Science (Physik, Informatik, Informationstechnik). Heidelberg, March 15-16, 1999, 67-86.*

Hilberg, W. (2000a). *Große Herausforderungen in der Informationstechnik. Vom Abenteuer der Forschung.* Groß-Bieberau: Verlag Sprache und Technik.

Hilberg, W. (2000b). Netzwerke maximaler Entropie. *Frequenz 54, 3-4, 80-86.*

Hilberg, W. (2002). Wie wirklich ist ein Gedanke? – Wittgenstein und die Informationstechnik. *Thema Forschung: Bionik, 2/2002, 04-109.*

König, W. (1994). *dtv-Atlas zur deutschen Sprache*. München: Deutscher Taschenbuch Verlag.

Li, W. (1992). Random Texts Exhibit Zipf´s-Law-Like Word Frequency Distribution. *IEEE Transactions on Information Theory 38, No.6, 1842-1845.*

Li, W. (2002). *Zipf´s Law*. http://linkage.rockefeller.edu/wli/zipf/

Mandelbrot, B. (1953). An information theory of the statistical structure of language. In: Jackson, W. (ed.), *Communication Theory: 486-502*. London.

Mandelbrot, B. (1959). A note on a class of skew distribution functions: Analysis and critique of a paper by H.A. Simon. *Information and Control 2, 90-99.*

Mandelbrot, B. (1961a). Final note on a class of skew distribution functions: Analysis and critique of a model due to H.A. Simon. *Information and Control 4, 198-216.*

Mandelbrot, B. (1961b). Post scriptum to "Final Note". *Information and Control 4, 300-304.*

Manning, Chr., Schütze, H. (2001). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: The MIT Press.

Meyer, J. (1989). *Die Verwendung hierarchisch strukturierter Sprachnetzwerke zur redundanzarmen Codierung von Texten.* Diss. THD.

Nachtwey, V. (1995). Textkompression auf der Basis von Wortnetzwerken und Grammatikmodellen. Diss. THD.

Rapoport A. (1982). Zipf´s Law Re-Visited. In: Guiter, H., Arapov, M.V. (eds.), *Studies on Zipf´s Law: 1-28*. Bochum: Verlag Brockmeyer.

Ries, Th. (2001). Über Möglichkeiten einer maschinellen Nacherzählung mit Hilfe eines hierarchischen Systems aus Sprachnetzwerken. Diss. TUD.

Shannon, C.E. (1951). Prediction and Entropy of Printed English. *The Bell System Technical Journal, Jan. 1951, 50-64.*

Siedenburg, K. (1992). *Symmetrische Netze für Parallelrechner*. Diss. THD. (VDI Fortschritt-Bericht Nr. 225).

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika 42, 435-440.*

Simon, H.A. (1960). Some further notes on a class of skew distribution functions. *Information and Control 3, 80-88.*

Simon, H.A. (1961a). Reply to "Final Note" by Benoit Mandelbrot. *Information and Control 4, 217-223.*

Simon, H.A. (1961b). Reply to Dr. Mandelbrot´s post scriptum. *Information and Control 4, 305-308.*

Simon, H.A. (1963). Some Monte Carlo estimates of the Yule distribution. *Behavioral Science 8, 203-210.*

Steinmann, F.-M. (1996). Netzwerkmodellierung und Segmentierung von Texten sowie Anwendungen zur Informationsverdichtung. Diss. THD.

Thompson, R.F. (1994). *Das Gehirn. Von der Nervenzelle zur Verhaltenssteuerung*. Heidelberg: Spektrum Akademischer Verlag Heidelberg.

Wittgenstein, L. (1914-1916). *Tractatus logico philosophicus*. Frankfurt: Suhrkamp.(1989)

Wittgenstein, L. (2001). *Philosophische Bemerkungen*. Berlin: Springer Verlag.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* Reading, Mass.: Addison Wesley, 1949 (Hafner Publishing Comp. New York 1972).

Glottometrics 5, 2002, 51-61 *To honor G.K. Zipf*

Power Law Models in Linguistics: Hungarian

Reinhard Köhler[1]

Abstract. First, the status of Zipf(-Menzerath)'s Law and its criticisms are discussed, and the application of power law models, particularly in linguistics, is supported from a general point of view. In the following sections, empirical studies on dependencies are conducted which test the Zipf-Mandelbrot Law, other power law models (Menzerath-Altmann's Law, the length-frequency dependency), and the word length distribution on data from Hungarian (a text and a dictionary).

Keywords: Zipf-Mandelbrot Law, Menzerath-Altmann Law, Power Law Models, word length, word frequency, Waring distribution, Conway-Maxwell-Poisson distribution, modified binomial distribution, length-frequency dependency, Hungarian

1. A General Remark on Power Law Models and on Certain Criticisms of Zipf's Law

Following a stochastic regularity in time, considerations can be found in the quantitative-linguistic literature which are sceptical about the lawfulness of the phenomena observed with rank-frequency distributions of words and other linguistic units. Every now and then, authors claim that these observations may not be considered as being generated by a meaningful linguistic law (the first one was probably Miller (1951, 91; 1957); the most prominent opponent to Zipf's work in general was Herdan (e.g., 1966)). The reasons given for their assessments vary to some degree but they all resemble each other in that they are of the type "Since a Poisson distribution of spaces in a string of letters produces the same Zipf-like distribution of 'words' as can be observed in real texts there is no point in assuming a linguistic law in this phenomenon". Others argue that the ubiquity of power law distributions (covering phenomena as divers as the rank-size distributions of words in texts, books of a given author in a library, salaries in a human population, size distributions of avalanches in sand-piles, of earthquakes and high waters, solar flares and pulsar glitches, catastrophes of any type etc.) contradicts their interpretation as a linguistically interesting fact in general, whereas in natural sciences a view is common according to which the study of power laws opens up new vistas (cf. Schroeder (1991) or even the ultimate understanding of the world (cf. Bak 1996).

We will here discuss the opinion that Zipf's law is uninteresting because there are mechanisms which yield, by simple mechanical procedures, similar (or even identical) results as the frequency count of words in texts do. Let us look at another example of mathematical modelling in quantitative linguistics, viz. certain distributions (such as the binomial, hypergeometric or others), which are often used as models of the distribution of properties of linguistic entities in texts. Modelling is simplification, which means in this case that by setting up statistical distributions as models of text properties, only very few properties (only the ones which are considered relevant for the purpose of the given investigation) are taken into account. The potential mistake of confusing the model with its original (i.e., reality) must be

[1] Address correspondence to: Reinhard Köhler, Fachbereich II – LDV, Universität Trier, D – 54286.
E-mail: koehler@uni-trier.de

avoided, i.e. the model does by no means represent the assumption that only the properties used for it would count. A quantitative examination of, say the probability of words to occur in a text block using the binomial distribution must not be interpreted as representing the assumption that a text is really generated by random processes such as taking words from an urn, and that grammatical, semantic, and pragmatic factors do not play any role in these processes. Or, if it turns out that the probability of the verses of some poems under analysis to begin with the phoneme /k/ follows the Poisson distribution, nobody would draw from this the conclusion that the poems were created by means of random numbers. If a law can be formulated in cases like the presented ones, these can be only *phenomenological* ones (i.e., laws which predict the data on the basis of a universal statement but do not reveal the underlying mechanisms; cf. Bunge).

On the other hand, if a mathematical model is a deduced consequence of a theoretically derived hypothesis, it can be used (under certain conditions) to formulate a *representational* law (i.e., one which not only predicts the behaviour of the system under consideration but also offers insights into its relevant mechanisms).

A model that is justified by theoretical considerations such as Mandelbrot’s (1953) derivation of his famous formula by setting up a hypothesis about language as a self-optimising system must be tested both theoretically (by checking plausibility and compatibility) and empirically by confronting the predictions of the model to data from reality.

However, the observation that there are other mechanisms which can produce similar data is of little concern here. Moreover, also a mathematical proof that a formula which represents a linguistic hypothesis can be regarded as the result of another (even a trivial) approach does not disprove the original hypothesis.

In most cases, however, where linguists or mathematicians publish their scepticism about the validity of Zipf’s Law the main argument is that the rank-frequency distribution with its “typical” shape be the “artificial” result of re-arranging the data according to their frequency. This criticism may be justified with respect to some recent publications in physics, where sometimes power law curves are considered as identical even if one of them is approximately a straight line in log-log coordinates whereas the other one appears as near to linear in linear-log coordinates (cf. e.g. Bak 1996). As opposed to that sloppy kind of analysis, in linguistics, the phenomenon has been scrutinised over and over again with the result that the rank-frequency distributions (and spectra) subsumed under Zipf’s Law show a very characteristic behaviour, which go beyond the simple fact that they are monotonously decreasing curves. There are, of course, infinitely many forms of curves that could yield from ranking data according to their frequency, but word frequency in texts follows the Zipf-Mandelbrot type of curve with surprising constancy and uniformity.

A closer look at the rank-frequency curves found with linguistic units reveals a specific property, which was detected and described by V. Kromer (1997): a typical indentation in the first part from top. Kromer was even able to find an explanation for this phenomenon, which can be considered as characteristic only of linguistic data. Thus, Zipf’s Law in the narrower sense is not only no triviality but also a very specific case among the varying forms of power law phenomena and deserves special attention and specific modelling.

2. Zipf-Mandelbrot’s Law in a Hungarian Text

The present study is in the first line based on dictionary data. However, a contribution to a Festschrift dedicated to Zipf should, if possible, contain a rank-frequency analysis of words in a text. Therefore, this empirical investigation starts with a Hungarian text (the short version of a diploma thesis in computer science with 2310 types and 5266 tokens). As can be seen from the following results[2], the rank-frequency distribution confirms to Zipf-Mandelbrot's Law and the spectrum to the Waring distribution, even if the diagrams show that the text is not a perfect example of the expected word frequency structure.

```
Distribution: Zipf-Mandelbrot (a,b; n = x-max)
Sample size: 5266
Parameters:
a = 0,831644926052134    b = 0,554612181736859    n = 2310
DF =1688
X² =798,4252     P(X²) =  0,0000     C =  0,1516
```

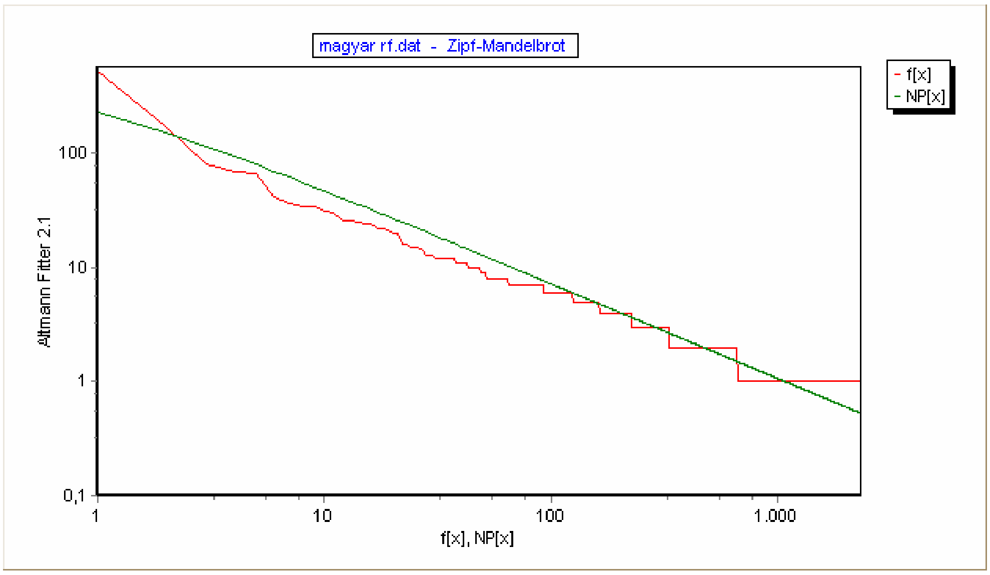


Figure 1: Fitting the Zipf-Mandelbrot distribution to the rank-frequency data

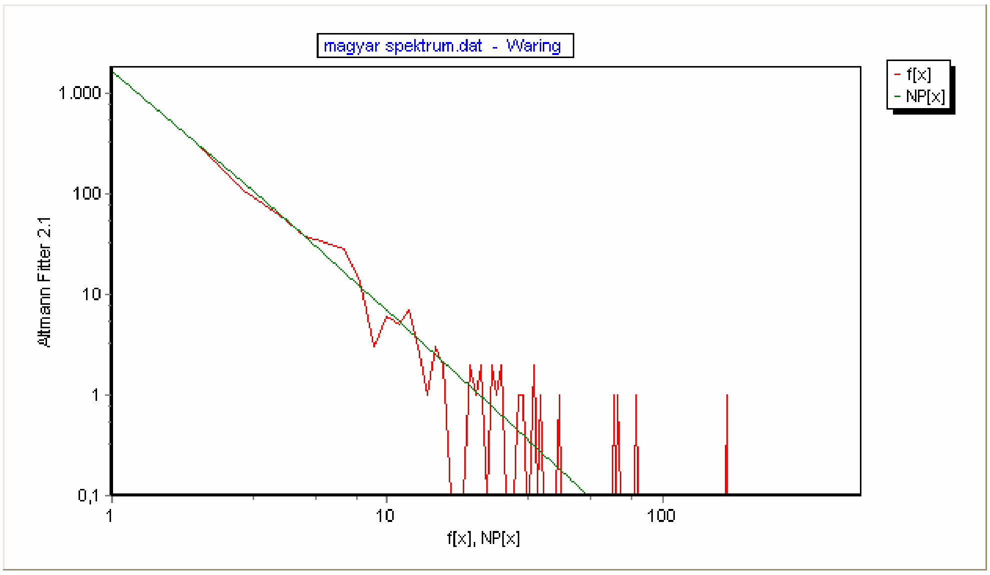


Figure 2: Fitting the Waring distribution to the spectral data

```
Distribution: Waring (b,n)
Sample size: 2310
Parameters:
b = 1,56185562623533    n = 0,626068264973653
DF =29
X² = 33,2356     P(X²) =  0,2683     C =  0,0144
```

[2] To fit theoretical distributions to empirical data, the Altmann Fitter (2.1) was used.

3. The Menzerath-Altmann Law and Word Length in Hungarian

Another kind of power law model frequently investigated in linguistics is the Menzerath-Altmann Law (MAL), which is represented by the formula

$$y = Ax^{B}e^{Cx} \tag{1}$$

(cf. Altmann 1980, Altmann, Schwibbe 1989), where y is the length of a constituent (e.g., mean syllable length), x the length of the corresponding unit measured in number of their constituents (e.g., word length measured in number of syllables). Formula (1) is a monotonously decreasing function, hence parameter B is a negative number.

This model has been tested on data of many languages, many text types and on each level of linguistic analysis – from sound length to sentence length, and even on a supra-sentence level (cf. Hřebíček 1995), and was confirmed in every case (cf., however, Rukk 2003, where Russian sentence length has been studied with a different result).

The present study contributes to testing this law on data from Hungarian, a language that has not yet been investigated in this way. Hungarian and other agglutinative languages are particularly interesting with respect to a hypothesis set up by Skalička (1966), who claimed that agglutinative languages must have longer words than other language types and syllables with a comparably low phonological complexity. This hypothesis seems plausible but has never been tested empirically up to now. In terms of the MAL, agglutinative languages should show flatter curves on the word-syllable level[3]. As this language type has not yet been investigated with respect to the MAL, a first study has been conducted here on the material of an electronic dictionary of Hungarian.

Hungarian orthography is rather phonological. Therefore, the graphematic representation of the words could be used with only little pre-processing (cf. below). Length of vowels and consonants has phonematic status in Hungarian, however, whereas vowels are marked by diacritics (such as "e" vs. "é", or "ü" vs. "ű"), length of consonants is marked by gemination ("t" vs. "tt"). As opposed to this point of view, some linguists think that Hungarian geminates should be interpreted biphonematical. For the present study, both alternatives have been taken into account – with apparently minimal difference in the results[4] (see below). Counting could be done automatically but some pre-processing was necessary to disambiguate cases as the following ones:

1. Letter combinations such as sz, cs, zs etc., which represent single phonemes (/s/, /tʃ/, and /ž/ resp.,) have to be disambiguated from identically looking sequences, which can occur at word boundaries within compounds.
2. Letter combinations which represent long versions of the above-mentioned consonants, such as in the compound noun "sasszem" (eagle's eye), where there is a word boundary between the letters s and sz, vs. "össze" (together), where there is none, or "mennyi" (how much) without a word boundary between n and ny vs. "ellennyugta" (counter-confirmation).

As can be seen from tables 1 and 2 and figures 3 and 4, function (1) could be fitted to the data with very good results.

[3] I owe this idea to Sabine Weber (personal communication).

[4] In fact, the differences are so small that a significance test can be considered as unnecessary.

Table 1

Syllable length as a function of word length in syllables.

Here, geminate consonants were counted as two phonemes. The single occurrence of a nine syllable word was excluded from the fitting because of its statistical unreliability.

Word length [no. of syllables]	**Mean syllable length**	**Expected syllable length**	**No. of words with length x**
1	3.279450	3.27065	1664
2	2.726930	2.75150	12733
3	2.530530	2.52829	20844
4	2.416290	2.40864	15620
5	2.350710	2.34053	5547
6	2.308170	2.30300	1510
7	2.288680	2.28573	289
8	2.270830	2.28299	60
9	2.222220	2.29129	1
A = 3.1432, B = -0.3067, C = 0.0398			
Proportion of variance explained (R^2) = 0.9987			

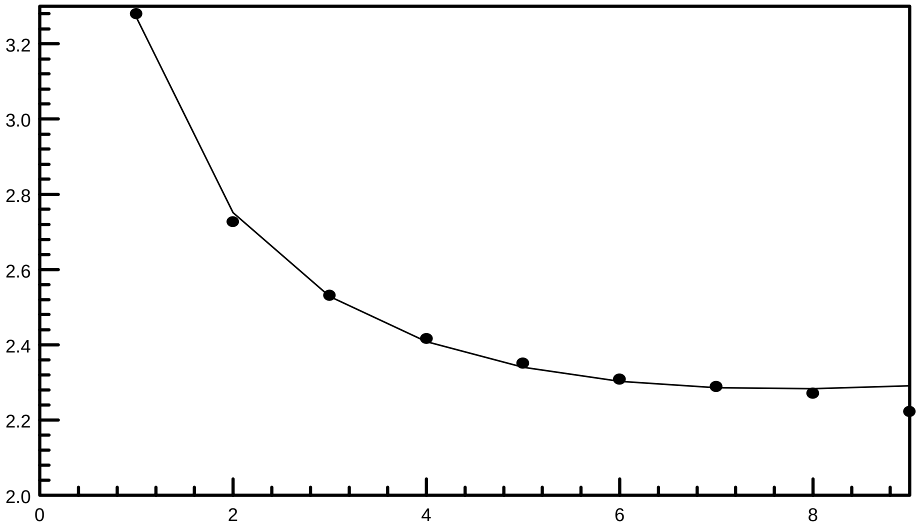


Figure 3: Graph representing the fit shown in Table 1

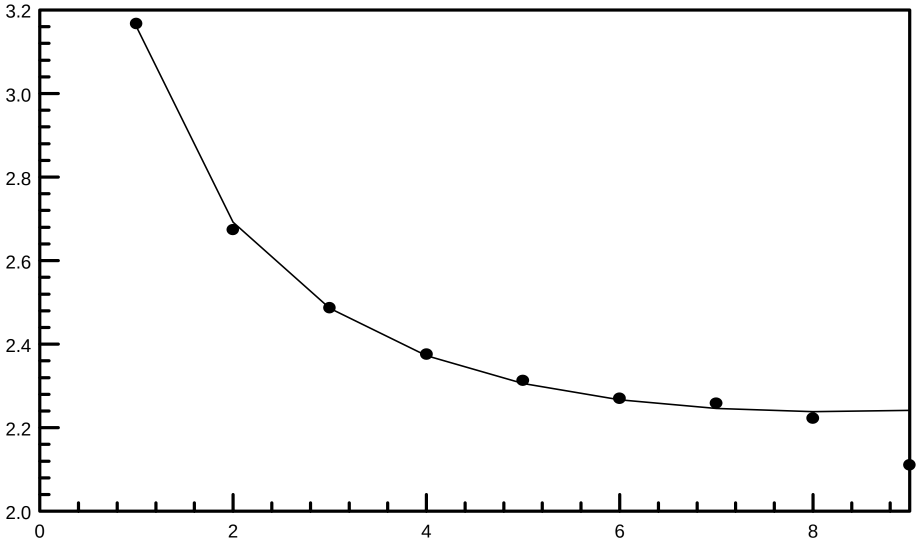


Figure 4: Graph representing the fit shown in Table 2

Table 2

Syllable length as a function of word length in syllables.

Here, geminate consonants were counted as long phonemes. The single occurrence of a nine syllable word was excluded from the fitting because of its statistical unreliability.

Word length x [no. of syllables]	Mean syllable length	Expected syllable length	No. of words with length x
1	3.167670	3.16069	1664
2	2.673640	2.69201	12733
3	2.487080	2.48577	20844
4	2.376170	2.37254	15620
5	2.312960	2.30588	5547
6	2.270310	2.26694	1510
7	2.258530	2.24636	289
8	2.222920	2.23890	60
9	2.111110	2.24135	1
A = 3.0545, B = -0.2808, C = 0.0342			
Proportion of variance explained (R^2) = 0.9988			

The results obtained here once more confirm the MAL on material from a language not yet evaluated so far.

4. The Menzerath-Altmann Law and Syllable Length in Hungarian

The Hungarian dictionary data allow for another MAL analysis, viz. on the syllable-sound level, however indirectly. As sound length was not available in terms of duration, the following approximation was used: Short vowels and consonants were counted as one length unit each, long sounds as two length units. Thus, only these two values could be obtained as length measure of a sound, and consequently mean sound length is a real number in the interval [1,2]. Table 3 shows the empirical findings and the result of the fit of formula (1) to the data.

Table 3

Sound length as a function of syllable length in no. of sounds.

The single occurrence of a nine sound syllable was excluded from the fitting because of its statistical unreliability.

Syllable length x [no. of phonemes]	Mean sound length	Expected sound length	No. of words with x syllables
1	1.178000	1.17455	1664
2	1.142000	1.15219	12732
3	1.144000	1.14109	20845
4	1.139000	1.13452	15621
5	1.134000	1.13042	5547
6	1.124000	1.12785	1510
7	1.124000	1.12634	289
8	1.126000	1.12561	60
9	1.050000	1.12546	1
A = 1.1701, B = -0.0333, C = 0.0038			
Proportion of variance explained (R^2) = 0.9195			

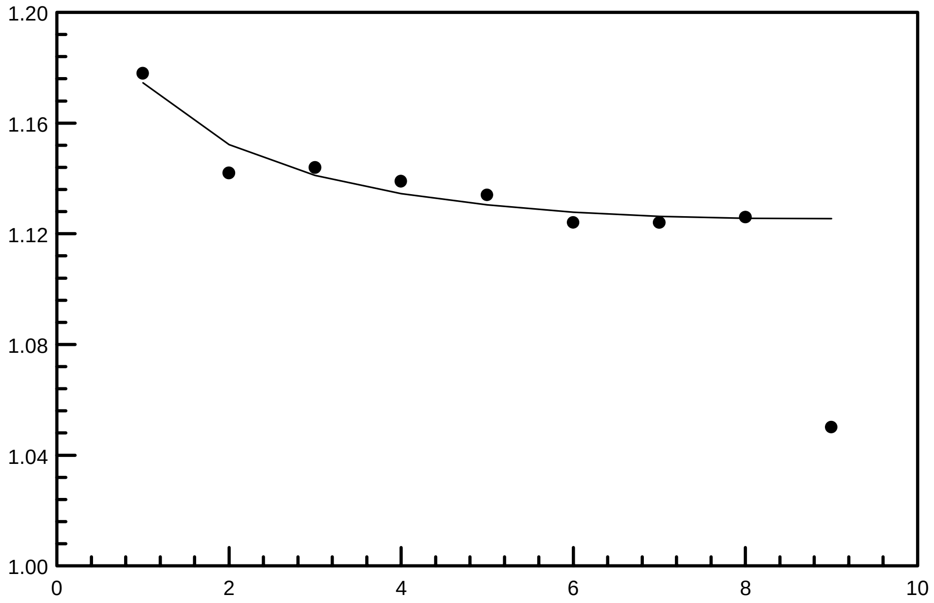


Figure 4: Graph representing the fit shown in Table 3

Dictionary entries without vowels (such as “b” – e.g., for the name of the letter *b* – or the horse command “brr”) were not taken into account.

The results presented in sections 3 and 4 are clearly another confirmation of the MAL. However, we cannot yet draw any conclusions with respect to Skalička’s hypothesis for the following reasons: (1) Sound length measurement in this study was only a rough approximation, (2) the results obtained here are based on a dictionary study whereas most of the available results of MAL studies were done on text data. Hence, a direct comparison is not possible, a test of significance is not applicable.

5. Word Length Distributions

According to Skalička’s hypothesis, agglutinative languages should have longer words than other languages. A possible test in this context is to compare the word length distributions of agglutinative to other languages, another one is to compare the differences between dictionary based length distributions and text based ones. Here, both types of length distributions have been studied.[5]

Word length in the dictionary is compatible with two models: the Conway-Maxwell-Poisson and the Hyperbinomial distributions.[6] Table 4 and Figure 5 show the fit of the Conway-Maxwell-Poisson distribution to the dictionary data.

The empirical distribution in the diploma thesis is shown in Table 5. Words with zero-length are, e.g. “s” – the short version of “és” (and) – and abbreviations such as “stb” (etc.). Within the framework of the Göttingen project (cf. Best 2001), abbreviations are expanded. In the present study, they were counted as they are.

[5] In general, word length distributions are studied on text material only (cf. Best 1997; Best 2001).

[6] For detailed information on discrete probability distributions see Wimmer, Altmann 1999.

Table 4
Fitting the Conway-Maxwell-Poisson distribution to the dictionary data

X[i]	F[i]	NP[i]
1	1664	1734,0768
2	12733	12500,8360
3	20844	21079,7406
4	15620	15195,4157
5	5547	5993,6900
6	1510	1481,0073
7	289	249,7238
8	60	30,4821
9	1	3,0275
a = 7,2089, b = 2,09595, DF = 6, X² = 91,6206 P(X²) = 0,0000 C = 0,0016 Sample size: 58268		

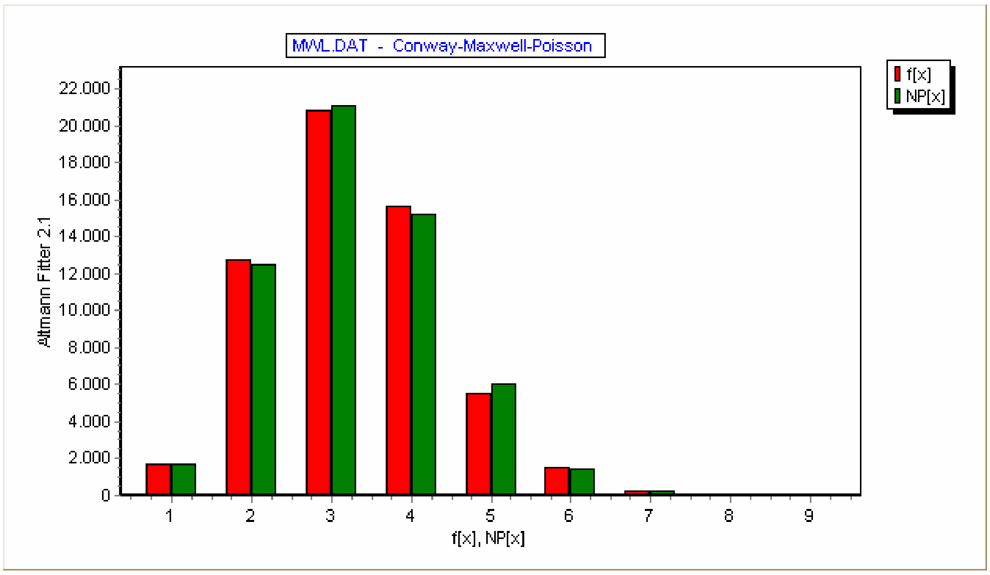


Figure 5: Graph of the fit shown in Table 4.

In analogy to the decision in Bartens, Zöbelin (1997, 196), the zero-syllabic words were added to the monosyllabic ones for fitting a model to the data. As the monosyllabic words show a specific behaviour, a modified distribution must be used (cf. Wimmer, Witkovský, Altmann 1999). In the present case, the Modified Binomial distribution could be fitted with good results (cf. Table 6 and Figure 6).

Table 5
The empirical word length distribution in the diploma thesis

Word length x [no. of syllables]	No. of words with length x
0	27
1	1544
2	979
3	1049
4	829
5	433
6	134
7	33
8	7
9	3

Table 6
Fitting the Modified Binomial distribution to the text data

X[i]	F[i]	NP[i]
1	1571	1568,9028
2	979	974,5889
3	1049	1099,0040
4	829	784,8911
5	433	398,2902
6	134	152,7061
7	33	45,9201
8	7	11,0962
9	3	2,6007
n = 21, p = 0,1013, a = 0,7703, DF = 5, X² = 15,30 P(X²) = 0,0091, C = 0,0030, Sample size: 5038		

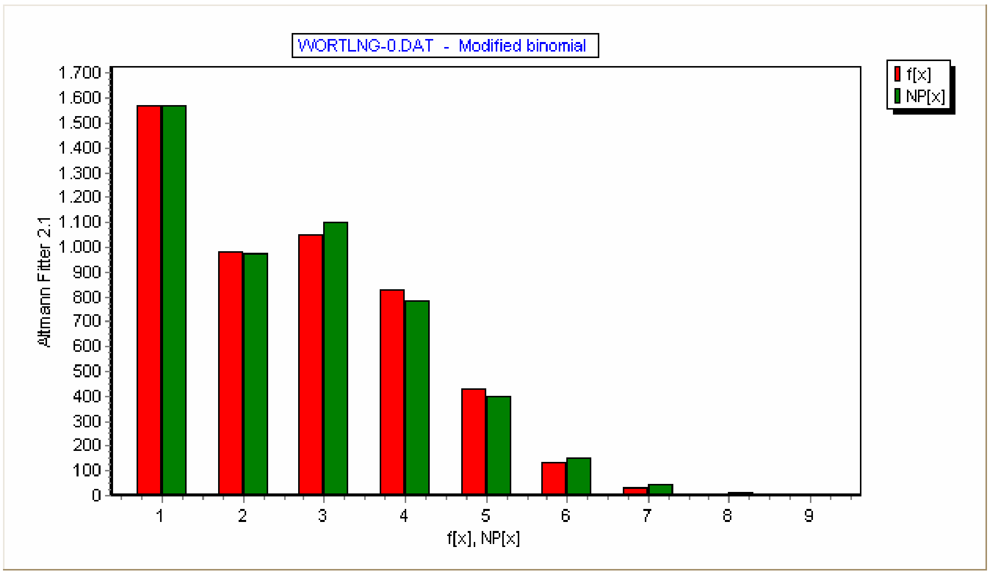


Figure 6: Graph of the fit shown in Table 6.

6. Word Length and Word Frequency

Yet another power law dependency Zipf was interested in is the function between word frequency and word length. Therefore, a short study on this dependency was conducted on the data of the text under analysis.

Here, a more general form of the power law is considered, viz. formula (2):

(2) $y = Ax^B$

where *y* represents mean word length, *x* is the frequency, and A and B are parameters. Figure 7 shows the fit of formula (2) to the data from the text. The parameters obtained are

A = 4.02732366
B = -0.332778471
and R^2 = 0.7580.

Whether the length variable shows an oscillation (cf. Köhler 1986, 137ff) in the frequency dimension must be checked on more Hungarian texts.

7. Conclusion

In the present study, first results of MAL studies on Hungarian data were obtained. This kind of results can be used, as well as the word length studies, to investigate particularities of agglutinative languages, especially to test Skalička's hypothesis. However, before any evaluation can take place, more Hungarian texts should be processed, and more agglutinative languages should be analysed.

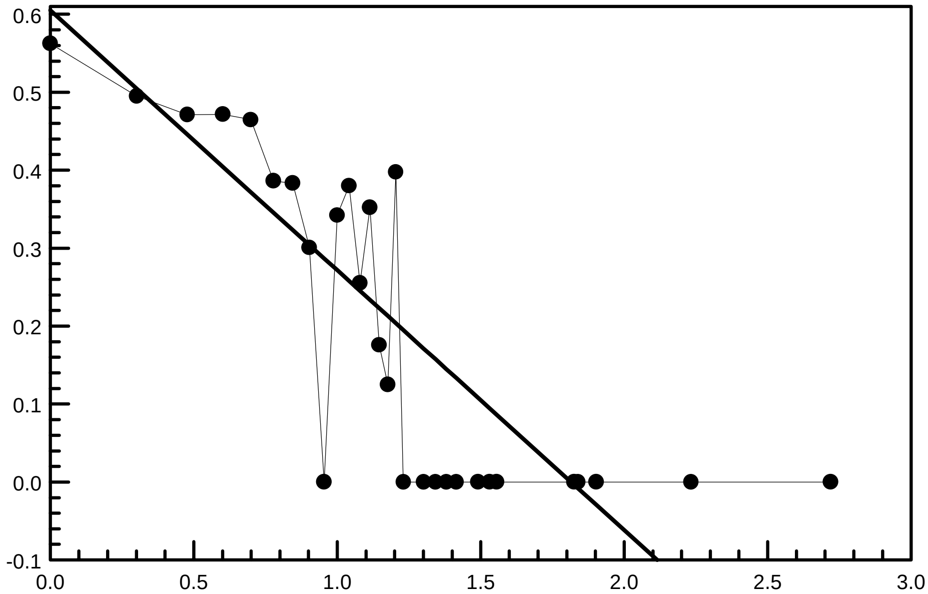


Figure 7: Length as a function of frequency: fitting the function $y = Ax^B$ to the text data. Both axes logarithmic.

References

Altmann, Gabriel (1980). Prolegomena to Menzerath's law. In: Rüdiger Grotjahn (ed.), *Glottometrika 2, 1-10.* Bochum: Brockmeyer.

Altmann, Gabriel, Schwibbe, Michael H. (1989*). Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Bak, Per (1996). *How nature works*. New York: Copernicus.

Bartens, Hans-Hermann, Zöbelin, Thomas (1997). Wortlängenhäufigkeiten im Ungarischen. In: Best (ed.) 1997: *195-203.*

Best, Karl-Heinz (1997) (ed.). Glottometrika 16*: The Distribution of Word and Sentence Length.* Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (2001) (ed.). *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt Verlag.

Bunge, Mario (1967). *Scientific Research I.* Berlin: Springer.

Herdan, Gustav (1966). *The Advanced Theory of Language as Choice and Chance*. Berlin, Heidelberg, New York: Springer-Verlag.

Hřebíček, Luděk (1995). *Text levels: language constructs, constituents and the Menzerath-Altmann law.* Trier: Wissenschaftlicher Verlag Trier.

Köhler, Reinhard (1986). Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Bochum: Brockmeyer.

Kromer, Victor V. (1997). Jaderno-veernaja model' vertikal'nogo raspredelenija slov v russkom jazyke. Dep. v INION RAN No. 52458 ot 31.3.97. Novosibirsk.

Mandelbrot, Benoît (1953). An information theory of the statistical structure of language. In: W. Jackson (ed.), *Communication Theory*: *486-502*. London.

Miller, George A. (1951). *Language and Communication*. New York: McGraw-Hill.

Miller, George A. (1957). Some effects of intermittent silence. *American Journal of Psychology 70, 311-314.*

Rukk, Maria (2003). To appear in: *Journal of Quantitative Linguistics*.

Schroeder, M. (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* New York: Freeman.

Skalička, V. (1966). Konsonantenkombinationen und linguistische Typologie. In: *Travaux linguistiques de Prague 1: 111-114.*

Wimmer, Gejza, Altmann, Gabriel (1999). *Thesaurus of Univariate Discrete Probability Distributions.* Essen: Stamm.

Wimmer, Gejza, Witkovský, Viktor, Altmann, Gabriel (1999). Modification of Probability Distributions Applied to Word Length Research. *Journal of Quantitative Linguistics 6/3, 257-268.*

Glottometrics 5, 2002,62-80 *To honor G.K. Zipf*

Laws and Theories in Quantitative Linguistics

Peter Meyer[1]

Abstract. According to a widespread conception, quantitative linguistics will eventually be able to explain empirical quantitative findings (such as Zipf's Law) by deriving them from highly general stochastic linguistic 'laws' that are assumed to be part of a general theory of human language (cf. Best (1999) for a summary of possible theoretical positions). Due to their formal proximity to methods used in the so-called exact sciences, theoretical explanations of this kind are assumed to be superior to the supposedly descriptive-only approaches of linguistic structuralism and its successors. In this paper I shall try to argue that on close inspection such claims turn out to be highly problematic, both on linguistic and on science-theoretical grounds.

Keywords: Zipf, language laws, ceteris paribus laws, emergence, complexity theory, explanation, science theory, word length, Menzerath

0. Introduction

Quantitative linguistics (henceforth, QL) as understood in the following considerations is concerned with accounting for quantifiable and measurable linguistic phenomena in terms of mathematical models such as curves, probability distributions, time series and the like. The mathematical formulas employed are attributed the status of scientific laws to the extent that they are deducible from very general principles or 'axioms' and are thus firmly integrated into some nomological network. Inasmuch as investigations in QL fulfil these basic requirements, they are believed to be paradigm cases, indeed the very first of their kind in the whole history of linguistics, of empirical scientific theories in a narrow, science-theoretically justified sense, as opposed to purely descriptive and taxonomical approaches in the traditional, 'qualitative' branches of linguistics, the latter being therefore charged with failing to attain the high methodological standards of the natural sciences (Altmann 2002, Altmann, Lehfeldt 2002 provide sufficient documentation for these claims). Likewise, G. K. Zipf, one of the founders of the contemporary methodology of QL, has lately been advanced to being "the first language theoretician", even the Newton of a new kind of theoretical, namely, 'Zipfian' linguistics (Altmann 2002, 22; 25).

The pungent methodological criticism of 'traditional', 'qualitative' linguistic approaches advanced in the QL literature constitutes a good starting point for the considerations to follow, since the basic science-theoretical and linguistic assumptions on which most work in QL is founded are to be understood as a corollary of this rejection of qualitative and descriptive methods. The criticism in question is, as I would like to show, based on a fairly restrictive and questionable notion of scientific theory. Thus for Altmann (2002, 19), 'qualitative' accounts for linguistic phenomena necessarily remain at a "proto-scientific level" because they cannot "satisfy the claims of natural sciences". In remarks such as these a certain normative science-theoretical stance becomes apparent, a stance according to which theories

[1] Address correspondence to: Peter Meyer, Seminar für Slavische Philologie, Georg-August-Universität Göttingen, Humboldtallee 19, D-37073 Göttingen; e-mail: meyer-peter@gmx.de

that deserve to be called empirical and scientific have to be modeled following the example set by the natural sciences, or strictly speaking, by axiomatized fundamental physics only. This attitude seems to date back to at least the days of the Logical Positivists and may firmly be said to be superseded by a much more differentiated contemporary discussion, represented, e.g., in textbooks such as (Balzer 1997). From the ample literature on the subject, it is usually only Mario Bunge whose works are quoted in support. Bunge (1995, 3) demands that linguistics confine its methods to quantifiable magnitudes that are referred to by the laws of a scientific theory proper:

> Can every feature be quantitated, that is, turned into a magnitude? I submit that only one property is, with all certainty, intrinsically qualitative, namely existence. I also submit that in every other case quantitation depends exclusively on our ability and interest, so that in the face of failure to quantitate we should suspend judgement and encourage others to try.

All this amounts to is an ill-founded profession of faith[2] that is refuted not only by well-established scientific practice in the social sciences and in the humanities. Recent science-theoretical surveys abound with reconstructions of purely qualitative theories (see, e.g., Balzer 1997 or Balzer, Moulines, Sneed 1987). Linguistics is a field that provides some paradigm cases of successful scientific modeling that fully meet explanatory requirements. The reconstructive methodology of historical linguistics is a case in point, providing, at least to a certain extent, even the possibility of prediction. There is no good reason not to call, say, some historical grammar of the Indo-European languages such as the one summarized in Beekes (1995) a partial (and admittedly not fully explicit)[3] *theory* of the historical development and relationship between the languages in question. The hypothetico-deductive method, which is often seen as a cornerstone of modern science, has undoubtedly become a paramount methodological instrument of contemporary 'qualitative' linguistics, mostly due to the influence of Chomsky's writings whose theory of human language, at least in its modern principles-and-parameters guise, belongs among the most convincing examples of 'non-numerical' theories, notwithstanding logically independent quarrels about whether it is good or 'approximately true' theory.[4] Even any good traditional grammar or dictionary of a language is a theory of that language, however incomplete, implicit and embryonic. Many concepts of such language theories are, *pace* Bunge, *inherently* qualitative, presupposing a yes-no decision that cannot reasonably be made 'fuzzy' or otherwise turned into a quantitative magnitude. A certain lexeme or phoneme either appears in this or that utterance token or it does not appear. *Tertium non datur*: A lexeme cannot be said to appear in the utterance "to a degree of 70%", although it would of course be possible to say, e.g., that "70% of all speakers think that lexeme A occurs in this utterance". But even in such a quantitative judgment a class of discretely individuated lexemes (among them, lexeme A) is already inevitably presupposed. As we shall see in section 6, it is precisely (and, from the point of view of QL,

[2] Ironically, the only qualitative property Bunge accepts at all, existence, is not regarded as a property at all by the vast majority of philosophers since Kant.

[3] It is important to notice that there are fundamental limits to explicitness even in highly formalizable theories of the 'exact sciences'. To give but one example, according to one of the most eminent researchers in this field of the philosophy of science, a deductive axiomatization of experimental physics is impossible (Suppes 1998).

[4] The theory of language advanced by Chomsky and his followers tends to be misdescribed in QL writings as a mere formalism to express language-specific rules of grammar that have been found by inductive generalization. As a matter of fact, however, contemporary generative grammar takes as its axioms, amongst other things, the uniform initial state of an assumed human language faculty and *deductively derives* from these axioms, amongst other things, predictions as to the implicit knowledge of an adult native speaker under given experiential boundary conditions. This implicit knowledge can be *tested* using empirical methods. Hence, modern generative methodology fulfils all requirements for theoryhood typically formulated in QL. See Chomsky (2000) for an accessible recent presentation.

ironically) modern complexity theory and related developments in abstract evolution theory that provide some deep arguments in favor of the inevitability and irreducibility of qualitative, descriptive and functional accounts of certain complex systems.

Bunge's demand that all linguistic properties be quantifiable is an outgrowth of a certain general view on what characterizes a genuine scientific theory. A concise recent summary of this view can be found in (Altmann, Lehfeldt [to appear]). According to the authors, the notions of 'explanation' and 'law' are adequate only in case the fundamental statements of the theory in question are of a *dynamic* nature in an abstract, quantitative sense, that is, representable as difference or differential equations or as self-regulation schemata in some sort of systems theory. Once again, we are left with some sort of metatheoretical *credo* that leaves open why *this* should be the only *modus operandi* allowed in modern science, let alone linguistics, particularly as the immense complexity of social and neurophysiological processes that jointly underlie the dynamics of language make it seem rather implausible that this dynamics can be modeled in any *interesting* way in terms of, say, a bunch of differential or equilibrium equations. The *ultima ratio* behind the methodological propensities typical of QL proponents seems to be the irresistible attraction exerted by the now-fashionable scientific paradigms of chaos and complexity theory; see below for some critical evaluation. For the time being, the only thing that can justifiably be said with respect to the methodology of QL is that it is a way of looking at language that is *complementary* to traditional approaches and cannot, for this reason, be translated into the conceptual apparatus of the latter or *vice versa*.

In a number of publications (e.g., Altmann, Lehfeldt 2002) the notion of 'theory' is defined in contradistinction to mere inductive generalization as allegedly offered by traditional, 'qualitative' approaches. To begin with, it must be stressed again that this allegation ignores the role of deductive-nomological explanations and of the hypothetico-deductive method in contemporary linguistics. Many of Chomsky's writings sound remarkably similar to recent contributions to QL in rejecting arbitrary inventorization of data and superficial empirical generalizations in favor of deep and unified explanatory principles from which empirical generalizations can be deduced; for a succinct early statement cf. Chomsky (1978). In this paper, I will assess several criteria proposed in the QL literature that are used to justify the status of a scientific theory for QL models. These criteria may be summarized as follows. The backbone of any scientific theory proper is formed by *laws*. Typical empirical hypotheses of QL (such as, say, distribution of word length in texts) are indeed laws in a strict science-theoretical sense since they are embedded in a nomological network, i.e. they are *deducible* from underlying postulates or 'axioms'; and, by virtue of referring to measurable quantities, these laws are subject to empirical *confirmation* or disconfirmation.

It is the main objective of this paper to maintain that the quantitative regularities discovered so far in QL do *not* pass as law-like statements, that is, as analogues to what qualifies as 'laws' in the natural sciences, particularly in fundamental physics. As a consequence, *explanations* for these regularities (in a science-theoretically established sense of 'explanation') are still wanting. It is, however, a delicate task to assess the possible impact and importance of the claims just put forth on QL research work. What is not claimed here is that the results obtained so far in QL are empirically or theoretically void. Nor will it be the purpose of the following remarks to impose certain normative science-theoretical requirements on QL. Quite to the contrary I would like to suggest that QL faces the danger of being caught up in a false picture of what constitutes 'real' science, a picture that could serve as a problematic guideline to further research work in drawing the wrong dividing line between what should be considered 'good' and 'bad' questions in QL.

Most of the critical reflections put forward in the present article are not novel. As early as in 1959, a stimulating review article on the book *Logique, langage et théorie de l'information* by B. Mandelbrot, L. Apostel, and A. Morf (Lees 1959) succinctly put its fingers on many of the conceptual problems and inherent limitations of a quantitative treatment of language. I shall permit myself to quote some of Lees's still relevant remarks in footnotes. In addition, it must be emphasized that some of the chief architects of contemporary QL are very explicitly aware of unresolved theoretical problems in the discipline (see esp. Grotjahn, Altmann 1993; Altmann 1999; Altmann 2002).

1. A case study: word length

For the purposes of our discussion I will take the theoretical treatment of word length distribution in German texts presented in Altmann, Best (1996) as a typical example of how a certain quantitative law-like statement can be taken to explain certain statistical regularities. Similar examples would of course have to be discussed for other putative laws of QL. I will simply assume here that the foundational problems observed with respect to my example case also arise in the context of other would-be QL laws, for analogous reasons.

The fact that the negative binomial distribution can be fitted well to the empirical distribution of word length (measured as number of syllables) in a wide variety of German texts is explained in this paper by stipulating an underlying self-regulative mechanism that consists in mutual influence between neighboring word length classes. In accordance with the general framework proposed in Wimmer et al. (1994), it is assumed that this influence implies a proportionality relation between neighboring classes:

(1) $$P_x = g(x)P_{x-1}$$

(1) is assumed to be the underlying *law-like principle*[5] that governs word length distribution in general. *g(x)* is a language-specific proportionality function that, for the German texts in question, is assumed to be representable as

(2) $$g(x) = \frac{a+bx}{cx}.$$

Here, *a* is taken to represent something like the length-invariant part of the German lexicon, whereas *b* is an author-specific modification factor (the author chooses to employ shorter or longer words, according to stylistic and other needs) and *c* 'stands for' the communicative interests of potential text recipients (such as minimizing the effort of decoding a given message). From (1) and (2), an explicit representation of P_x can be derived. The deduction yields a family of univariate discrete probability distributions with two parameters, namely, that of the negative binomial distribution. Now this family of distributions can indeed be fitted to the German text data in question. Thus, the underlying principles seem to receive empirical confirmation. In what follows, we will examine in turn the different parts of the methodology just sketched.

[5] Thus, Wimmer, Altmann (1994) write: "We consider statement (1) as a law-like hypothesis since it fulfills the requirements put on laws …, above all generality, systemicity and confirmation. Nevertheless it is merely a skeleton that must be filled with flesh taken from languages, genres and authors, all of them bringing different boundary and subsidiary conditions which can vary in the life of language or of an author. Thus no *specific* formula following from (1) holds eternally for all languages or even one language."

2. The notion of law

It will be useful to compile some definitional statements about ‘lawlikeliness’ as found in the QL literature. Altmann (1993) writes, quoting Bunge in support:

> “Only syntactically well-formed, semantically meaningful general statements that are empirically testable, not including observational concepts, stating something about invariances and going beyond our present knowledge should be considered as hypotheses. If a hypothesis is derived from assumptions (axioms) or from a theory, if it is corroborated by an empirical test and if it can be connected with other similar statements (systematized), then we can call it a law.”

In what follows I will examine in particular whether QL hypotheses are indeed derivable from assumptions or axioms (section 3) and to what extent they can be corroborated empirically (section 4).

Empirical confirmation of QL hypotheses such as (1) above is impeded by the possibility of ‘exceptions’ that have to be taken account of in some way. Altmann, Erat, Hřebíček 1996 write on behalf of the empirical validity of (1):

> We can consider the probability distribution as a kind of attractor, a form existing in every language – i.e. existing unconsciously in the text users – to which the empirical distributions of the given variable tend. Of course, there can be a number of different attractors exerting their impact on individual writers or on individual genres, and each of them can evolve in the course of time. As a matter of fact, formula (1) merely represents a mechanism which can take into account a number of boundary or subsidiary conditions, as is the case in natural laws, too. In practice, if a text deviates from the supposed attractor, we say that it wanders to another attractor, which is quite a normal circumstance in the life of a text producer. This wandering can be expressed in different ways (cf. Wimmer et al. 1994), e.g. in the modification of *g(x)*, in the increase of the order of the difference equation (1), in the modification of the individual frequency classes, in the mixing, compounding or convolution of probability distributions, etc.

Section 4 will also criticize the strategy hinted at in this quotation, namely of attributing QL hypotheses the status of *ceteris paribus* laws that are effective only under certain boundary conditions that cannot be listed explicitly and exhaustively.

In addition, the above quote also refers to underlying ‘mechanisms’ that are supposed to be described by the laws of a theory.[6] Section 5 will deal with some of the stipulated language mechanisms that are assumed to ‘generate’ the quantitative regularities that have been observed so far.

3. Deriving laws from axioms

Turning back to our case study, it must be stressed at the outset that a purely mathematical deduction of the negative binomial distribution from the difference equation (1) plus specification (2) does not provide us with a theoretical explanation of the distribution, since it does not embed it in a nomological network that has an *independent justification*. (1) and (2) are nothing but a mathematically equivalent reformulation of the probability distribution.[7]

[6] Thus, Altmann (1993) writes: “**Laws** are statements about mechanisms which generate observable phenomena.”

[7] Altmann (1980) acknowledges this point, when he comments on the derivation of Menzerath’s law from a simple differential equation: “The derivation from a differential equation is not sufficient in order to award the statement (4) [Menzerath’s law in a quantitative formulation, P.M.] the status of a law. It remains a theoretically not fully validated hypothesis as long as it is not set in relation to other laws of language, i.e. until it is

What we need is a testable criterion that tells us when it is appropriate to assume that (1) and (2) hold.

Recent efforts in QL (cf. Wimmer, Altmann 2003) concentrate on finding a 'unified derivation' of linguistic laws. This amounts to finding a very general class of difference / differential equations[8] from which all those formulas that have been employed in descriptive QL models so far can be derived. While the principle idea of the authors consists in epistemically integrating disparate research domains under the heading of a new 'supertheory' that contains the old particular ones as special cases, we are eventually left with the observation that different mathematical *formulas* employed in QL – representing variously probability masses or densities or function curves – may be transformed into some sort of very general 'canonical form'. There are no good reasons to assume that this purely formal analogy between extremely different formulas used in wildly disparate *interpretations* (*as* probability densities, *as* functions etc.) has a deeper reason connected somehow with (universal) properties of *human language*. All we get is a purely mathematical observation that has not yet any clear implications for the phenomena described with the aid of the respective formulas.

To sum up: The observation that the family of probability distributions defined by (1) and (2) can be used to 'model' word length in a variety of texts does not make (1) (or, for that matter, (1) *cum* (2)) a law statement, let alone a deductive-nomological explanation of observed word length distributions. For this to be the case, we would need some further justification for positing something like (1) as the general principle governing word length in human language. Interestingly, Wimmer et al. (1994) hint at some ideas to give (1) some initial plausibility:

> We assume that the various word length classes do not evolve independently of each other. If there is a gradual increase of disyllabic words in a language evolving from monosyllabism to polysyllabism …, this occurs as a function of the number of new meanings that must be coded and of the degree of polysemy and redundancy in the class of monosyllabic words. If the redundancy in this class has reached a critical level, the equilibrium must be restored by means of functional equivalents, e.g. tones, new phonemes, extension of phoneme distribution, variation of word length, etc. If a language has recourse to polysyllabism, the class of disyllabic words must necessarily be made proportional to that of monosyllables, i.e. in probabilistic terms
>
> $$P_2 \sim P_1.$$
>
> If a language is no longer restricted to monosyllabism and polysyllabic words are introduced, then self-regulation comes into play and controls the whole frequency structure of word length.
> In the first step, i.e. when disyllabic words are introduced, proportionality can be considered as constant, i.e. $P_2 = aP_1$. If longer words come into existence, constant proportionality will be replaced by a function of length $g(x)$. We thus obtain the basic formula
>
> $$(1) \quad P_x = g(x)P_{x-1}.$$

Even for the simple case of a 'monosyllabic' language developing disyllabic words, however, the authors' argument is far from clear. For a given text in the language in question, two probabilities must be posited, namely, the probabilities P_1 and P_2 that a given word form token in the text consists of one and two syllables respectively, where $P_1+P_2=1$, hence the

incorporated into a system of laws. Such a system does not exist at present, we merely suspect that it is somehow connected with the principle of least effort or with some not yet known principle of balance recompensating lengthening on one hand with shortening on the other."

[8] The approach runs roughly as follows: The relative rate of change of a variable Y is taken to be dependent on the rate of change of one other independent variable X. The latter is itself controlled by different powers of X that are associated with different multiplicative factors.

proportionality factor *a* is already given as $a = \frac{1}{P_1} - 1$. Now it is obvious that one cannot say of two *numerical values* that they stand in a proportionality relation to each others since only *functions* can be proportional to each other (*f(x)* and *g(x)* are proportional to each other iff *g(x) = const. · f(x)*). Thus, the proportionality posited by Wimmer et al. makes sense only inasmuch as word length probability is considered a function of some independent variable, e.g., text length. Empirical results show that the proportionality factor will vary from text to text, a point also stressed by the authors. Hence, the proportionality factor as assumed by the authors implies some kind of counterfactual conditional of the following kind: "Had we examined another text that is in all relevant respects similar to *this* text, we would have obtained the very same ratio of disyllabic to monosyllabic words." As long as we do not have any non-circular account of what the phrase "in all relevant respects" means here, the conditional is virtually devoid of meaning. And even if we had such an account we would still be in need of a scientifically valid justification for the quantitative principle (1) besides the air of vague qualitative plausibility. Hence, the 'proportionality principle' (1) remains unjustified in a strict sense of the word even for the simple case of a language restricted to two word length classes and *g(x) = const.*, since we cannot figure out what proportionality should mean here at all. Analogously, for the general case, in which we do not have a proportionality *constant* but a 'proportionality function', principle (1) becomes literally tautological because, trivially, for any discrete univariate probability distribution *P(x)*, (1) can be made true by defining a 'proportionality function' *g(x)* that fulfills the equation $g(x) = \frac{P(x)}{P(x-1)}$. It is clear that principle (1) gains empirical character only by virtue of specifying *g(x)*. However, we are not offered any theoretically well-founded restrictions on to what class of functions *g(x)* should belong, only some inductive evidence on what functions 'worked well' in past investigations, that is, have led to a good fit for a reasonable amount of texts. Nor do we possess any criterion for predicting which selection among a set of 'approved' functions will do well for a newly investigated text. In other words, (1) cannot be falsified; hence, it is not an empirical principle in any sense and, therefore, is not capable of being an 'axiom' from which theorems of word length distribution could be derived.

The reader might wonder whether the preceding remarks do not miss the very point of theoretical reasoning in QL since it is wrong to require that the 'principles' from which we deduce testable theorems be themselves deducible or justifiable in terms of yet other principles or laws. Naturally, or so QL adherents may argue, at some point deduction must come to an end and we arrive at the 'first principles', that is, the 'axioms', of the theory, for which further justification is neither possible nor required. The extremely successful foundational equations of, say, Newtonian or quantum mechanics are far from intuitive plausibility (for the less obvious case of Newton's second law see Weizsäcker 1985), but of course nobody would reject them for this reason. But why should anyone wish to embrace Newton's second law as an axiom of classical mechanics but nevertheless deny some explicit version of (1) or 'Zipf's law' or 'Menzerath's law' the status of an axiom of QL, although in both cases empirical confirmation of statements (theorems) deduced from the alleged laws is indeed possible? The difference between the two cases is connected with the important fact, overlooked by classical Logical Empiricism, that not every formal deductive system is a theory, let alone an empirical theory. In the case of Newtonian mechanics we "know" which empirical (real-life) systems Newton's three laws should be applicable to, that is, we are able to specify, though in a necessarily pragmatic yet non-circular fashion, the so-called 'intended systems' of Newtonian mechanics. The specification of the set of intended systems is an essential constituent part of any empirical scientific theory, besides the formally defined 'models' or 'axioms' themselves.[9] The point is trivial enough to be restatable in any science-theoretical framework: We must have some criterion that tells us which empirical phenomena the theory 'talks about', makes predictions of etc. This criterion must be independent of the specification of the theory's axioms.[10] Otherwise, astrology would count among the respectable scientific theories for the sole reason that its applicability could be restricted *post hoc* to those cases where prediction turned out to be successful. As far as our principle (1) (where *g(x)* must be assumed to belong to a family of previously specified functions) is concerned, there is no clearly defined class of intended systems for it. All we can point to is statements about some families of probability distributions that 'work well' for word length in a vaguely specified range of texts. We have no idea why (1) – in a version specified for *g(x)* – cannot be applied to this or that text; prediction is impossible. So all we can safely say is that (1) holds whenever it is found to hold. In this respect at least, QL does not yet fare much better than astrology.

The problem of parameter interpretation looms large in the above example just as elsewhere (cf. Altmann 2002). The parameters that appear in the various 'proportionality functions' proposed so far in the literature suffer from a complete lack of interpretability; they are just numbers obtained by fitting the model function class to the data at hand and vary from text to text without being predictable or connectable to other empirical statements about the texts in question. To be sure, interpretations have been proposed but they are plainly *ad hoc* and not susceptible to any sort of confirmation. In the interpretation proposed by Wimmer et al. (1994), the parameters involved come to be loosely associated with Zipfian 'forces'. None of the three postulated 'factors' is measurable (and no method of measuring seems to be forthcoming either), so empirical confirmation of the interpretations is impossible. Note that the authors do not even have a proper justification for associating the three factors assumed with the three parameters in the way they actually do. Why should parameter *c* 'stand for' the communicative interests of communicants? Increasing *c* indeed shifts the probability distribution toward smaller average word length. This seems to be the motivation behind the interpretation associated with *c*, since short word lengths will be favoured on the production side of the communication process. But, of course, we might as well assume *a* or *b* to be the parameter associated with the factor in question, as long as we say that *a* (or *b*) is inversely proportional to the communicative interests of speech producers. We must conclude that the interpretations suggested so far are but a 'fifth wheel', an ornament that plays no empirically testable role within the theoretical apparatus at all.

- Note that parameter interpretability does not imply some obscure requirement to the effect that the parameters correspond to directly observable magnitudes. A (variable) parameter may be said to be interpretable just in case there is *another* theory or law that makes an *independent* statement about the values the parameter may have. In other words, the parameter must appear in at least *two logically independent law statements*: The 'nomological network' must be tight enough to avoid the danger of immunization against falsifiability.

[9] See Balzer, Moulines, Sneed (1987) for details concerning the concept of 'intended system' introduced here. Note that the point I want to make here does not hinge upon selecting the particular 'non-statement' or 'structuralist' science-theoretical framework as presented in these two books. The authors use the term 'intended applications' for what I, following Balzer (1997), prefer to call 'intended systems' here.

[10] The range of intended systems should not be thought of as a neatly pre-defined set. As science progresses, new candidates for intended systems may be discovered. In this case, the axioms of the theory indeed define necessary conditions for candidatehood, a phenomenon called '(partial) autodetermination' in the structuralist approach. This must not be taken to imply that, at least in some cases, no independent specification of a class of intended systems is necessary since autodetermination is a phenomenon of theory development through time. Synchronically, criteria for recognizing the currently agreed-upon intended systems of a theory must be sufficiently clear-cut to meet with consensus of the scientific community.

However, it is precisely this nomological network that is not yet in sight for QL. Lack of interpretable parameters turns out to be yet another indication for lack of theory.

4. Inductive corroboration (testability)

The stochastic hypotheses advanced in QL cannot be confirmed directly but only disconfirmed. Usually standard Neyman-Pearson hypothesis testing is adduced to show whether linguistic data, say, the distribution of word length in a corpus of texts, is compatible with a certain hypothesis. There are difficult problems with this type of statistical confirmation, many of which are discussed at length in Grotjahn, Altmann (1993). The main problem is that if fitting of the proposed probability distribution or curve is successful, we have only shown that there is no good empirical reason to *reject* the hypothesis: *If* the hypothesis were correct, *then* the observed discrepancy between the actual data and the expected values according to the hypothesis would not be too improbable. Usually, a large number of quantitative models can be applied successfully in the sense that fitting is possible. The difficult question is to select the 'right' model; it can only be solved using *theoretical* considerations that are still absent because we have no principled methodology for comparing different models or different 'derivations' of the same model.

In order to test "general hypotheses in which no observational concepts (such as sound, syllable, word) occur" (Altmann 1980), it is necessary to use 'observational concepts', that is, theoretical terms as used in qualitative descriptive linguistics. In investigations on word length, we need *operational criteria* that individuate words and tell us how long they are (as measured in syllables or phonemes). It is standard practice in QL to use whatever rough-and-ready criteria are at hand (take words to be sequences of letters between blanks etc.), as long as the procedure chosen leads to statistically significant results. In its contemporary guise, QL has no choice here, since there are no known 'derivations' of observed quantitative regularities that would somehow theoretically reflect the *nature* of the qualitative concepts presupposed, i.e., the *role* they have to play in the qualitative theories that give them their meanings. It is very difficult to see whether such derivations would be possible at all since QL is not an 'autonomous' linguistic subdiscipline in the sense of Itkonen (1983); it presupposes *with conceptual necessity* an 'autonomous', i.e. traditional and qualitative, description of linguistic utterances. It is important to understand why the stochastic methods of QL cannot *replace* a qualitative treatment of notions such as 'word' or 'syllable'. Standard linguistic notions are *token-based*, that is, they presuppose the possibility of deciding for any given *token utterance* of which words, syllables, etc. it consists. This decision hinges upon the theoretical role of the notions employed, that is, their role within the qualitative theory of that language. QL notions can only be used to make assertions about statistical populations of utterances and do thus not have the capacity of making theoretically relevant descriptive decisions on single token utterances. Hence, even if a certain qualitative concept of 'word in language L' does not lead to interesting quantitative generalizations about texts in L, this can no more be taken to imply that the qualitative concept is to be replaced by another, 'better' one, than any amount of neuro- or psycholinguistic research can lead to 'amending' or 'revising' a qualitative concept. Even if we had a theoretically deducible stochastic regularity that works well with a certain concept (set of criteria) C_1 of 'word' but does not work at all with another set of criteria C_2, this would not tell us that we should henceforth use C_1 instead of C_2 in our *qualitative* descriptions since the viability of a qualitative concept can only be judged relative to the qualitative theory it forms a part of. It is the qualitative delimitation of the concept that gives a stochastic statement its meaning in the first place. Observable statistical regularities about artificial constructs that have no independently discernible place in a linguistic description (as might be the case with C_1) are virtually meaningless. As B. Mandelbrot has pointed out, qualitative and stochastic treatments of language are mutually incompatible and complementary; I would like to add that the stochastic treatment is conceptually dependent on the qualitative one but not *vice versa*.

It should be clear from the outset that none of the quantitative 'laws' proposed by QL so far can be expected to hold without exceptions, in striking contradistinction to the laws of fundamental physics. Given any QL 'law', it is always possible to artificially construct a counterexample, say, a text violating the stipulated stochastic regularity. As a matter of fact, exceptions to the inductive generalizations proposed in QL work are found anyway as soon as sufficiently large corpora of samples (usually, texts) are examined. The observations just mentioned could, and indeed should, be taken as an indication that QL does not have the same science-theoretical architecture as, say, fundamental physics, where 'laws' are assumed to hold without exception. Generally speaking, there are no good *a priori* reasons to believe that any 'proper' scientific theory must have laws in the very same sense that a small subset of the natural sciences is based on laws. QL proponents, however, would like to see the scientific apparatus of QL in complete conformity with that of the natural sciences.

It is often assumed that the inductive generalizations of QL are indeed 'laws' proper, if only a special kind of them, to wit, *ceteris paribus* laws that hold only when certain necessary preconditions are satisfied. However, since those necessary preconditions can – in virtue of the *ceteris paribus* restriction – not be stated explicitly and are, therefore, not specified by the law itself, the *ceteris paribus* clauses amount to no more than a trivial immunization strategy.

Thus, when one assumes, with Altmann (2002, 22), that "language laws hold only for homogeneous data" then it is difficult to avoid downright circularity, as homogeneity of data can most likely be defined only in terms of the law in question: data are homogeneous just in case the law is applicable to them.

Moreover, in the case of supposed QL laws, violations can be produced systematically and intentionally, as I already pointed out. To give but one example, given any specific candidate for a word length distribution regularity in natural language texts, it is possible to systematically construct texts that cannot be subsumed under the proposed regularity. Regularities that can be violated in an operational manner are neither laws nor *ceteris paribus*-laws (see Mott 1992, 462 for further elaboration of this point).

The reader might protest here, pointing out that there does indeed exist a well-established notion of *ceteris paribus*-law in science theory. In a similar spirit, Lehfeldt, Altmann (2002, 331; 341) hint at the oft-repeated claim that laws in *all* sciences always come with some *ceteris paribus* clause, obviously to turn down the suspicion that QL might be a second class science, when compared to fundamental physics. However, the historical source of such opinions is a misunderstanding of an important insight by Hempel formulated, e.g., in Hempel (1988). Earman, Roberts (1999) provide an excellent discussion of this point. The following quotations may serve as a summary of their argument:

> Hempel's claim is that typically a theory T of the advanced sciences will not have *any* logically contingent consequence S whose non-logical vocabulary belongs entirely to V_A [the set of 'antecedently understood terms' of T, P.M.]. What we can hope to derive from T are consequences of the form $P \rightarrow S$, where again S is a logically contingent sentence whose non-logical vocabulary belongs entirely to V_A and P is a "proviso" that requires the use of V_c [the set of theoretical terms first introduced with T, P.M.].[11]
>
> Hempel's provisos are … simply conditions of application of a theory which is intended to state lawlike generalizations that hold *without* qualification. Indeed, Hempel makes it explicit that his provisos are clauses that must be attached to *applications of a theory* rather than to law-statements…[12]

[11] Earman, Roberts (1999, 442).

[12] Earman, Roberts (1999, 444).

There is, in fact, an ongoing debate in the philosophy of science about whether a substantial notion of empirically non-void *ceteris paribus* laws can be found at all. J. Earman and J. Roberts provide a careful and painstaking survey of recent proposals on saving *ceteris paribus* laws from vacuity and argue that "not only is there no persuasive analysis of the truth conditions for *ceteris paribus* laws, there is not even an acceptable account of how they are to be saved from triviality or how they are to be melded with standard scientific methodology" (1999, 439). It is worth quoting their conclusion at some length:

> There *is* a clear sense to be given to the notion of a "near-law", i.e. a generalization that is not a strict law, but that deserves to be called a "near-law" because it is, in a precise sense, true or approximately true in almost all intended applications, because it plays the role of laws in giving explanations, supporting counterfactuals etc., and because it is clear that it makes definite claims about the world and can be confirmed or disconfirmed empirically. But, we claim, the most clear paradigms of such laws (viz. the laws of phenomenological thermodynamics) are not thought of as *ceteris paribus* laws, and statements that are thought of as *ceteris paribus* laws do not answer to this clear sense of a "near-law". […]
>
> In the light of this, we wish to make the following suggestion. "*Ceteris paribus* laws" are not what many philosophers have taken them to be, that is, they are not elements of typical scientific theories that play the same kinds of roles in the practice of science that less problematic statements such as strict laws or near-laws (in the sense just defined) play. Rather, a "*ceteris paribus* law" is an element of a "work in progress", an embryonic theory on its way to being developed to the point where it makes definite claims about the world. […] To revive a now-unfashionable notion, "*ceteris paribus* laws" belong to the context of discovery rather than to the context of justification. […]
>
> If laws are needed for some purpose, then we maintain that only laws will do, and if "*ceteris paribus* laws" are the only things on offer, then what is needed is better science, and no amount of logical analysis on the part of philosophers will render the "*ceteris paribus* laws" capable of doing the job of laws (1999, 465-466).

Earman and Roberts concede that the remarks just quoted look "at first glance to be a negative judgment about the special sciences as compared with fundamental physics". However, their intent is to reject a "misguided egalitarianism about the sciences":

> It is not "*ceteris paribus* all the way down" – *ceteris paribus* stops at the level of fundamental physics. But we are *not* physics chauvinists […], for we deny that the mark of a good science is its similarity to fundamental physics. The concept of a law of nature seems to us to be an important one for understanding what physics is up to, but it is a misguided egalitarianism that insists that what goes for physics goes for all the sciences. The special sciences need not be in the business of stating laws of nature at all, and this blocks the inference from the legitimacy of these sciences to the legitimacy of *ceteris paribus* laws. For us, it is ironic that an effort to justify the special sciences takes the form of trying to force them into a straitjacket modeled on physics. We think this effort should be resisted, since it damages both our understanding of the special sciences and our understanding of the concept of a law of nature (1999, 472).

5. Mechanisms or metaphors?

In view of these difficulties it is natural to look for a *scientific* instead of a *science-theoretical* treatment of systems that show a certain behavior in a 'more often than not' fashion which is not open to a deterministic or mechanistic micro-level description. And indeed the past few decades have seen the rise of a whole bunch of scientific disciplines – theories of complexity, catastrophe, chaos, dissipative and self-organizing systems – that deal with phenomena of this kind. So it is hardly surprising to find many QL researchers using concepts like 'attractor', 'self-organized criticality' and 'synergetic order parameter' as background metaphors for quantitative descriptions of linguistic phenomena.

Of course, it is the received view of modern QL that these concepts are *not* used merely metaphorically: language is assumed to simply *be* a self-organizing system that functions in a way analogous to, say, self-regulating dissipative systems in chemistry. However, as Kanitscheider (1998, 23) emphasizes, the mere *claim* of analogy is not enough; it is a hypothesis that has to be *proved* on empirical grounds. Hence, transferring a formal model such as Haken's synergetics to a new domain of phenomena is tantamount to setting up a new *theory* that must be validated *independently* of previous applications of the model in other domains.[13] In other words, the right motto should be: first set up your linguistic theory, then try to find a common denominator with theories from other fields.

A typical example of an ill-defined formal metaphor is the rather overused notion of *attractor*. Take the following statement: "Drawing on chaos theory, one supposes that theoretical linguistic units – often referred to as "-emes" in theoretical language – are attractors, and that variations and changes in the unit represent shifts toward another attractor […]. Only against this background is it possible to speak of self-organization in language, anyway" (Altmann 1996). It may well be doubted whether the attractor concept, when applied to notions of traditional linguistics, provides any additional theoretical insights or whether all we get is a vague feeling of plausibility. If a statement like "phonemes are attractors" is to have any empirical content, we must have a sufficiently elaborated mathematical concept of 'attractorhood' that may be applied to linguistic data in order to generate empirically testable hypotheses. Nothing of this kind seems to be in sight today.

It might be instructive at this point to discuss a particular example in some depth in order to shed some light on the purported *explanatory power* of the attractor metaphor. Lehfeldt, Altmann (2002) try to account for a certain sound change in Old Russian, the so-called fall of the two yer vowels.[14] Their theoretical starting point is 'Menzerath's law' that, in its modern quantitative version (which is now usually called the Menzerath-Altmann Law; principal references are Altmann 1980 and Altmann, Schwibbe 1989), relates the length of a linguistic construction to the length of its constituents through an inverse proportionality, or, more generally speaking, through a power law. Applied to word length, the 'law' *in its basic form* may be written as the equation $Y = Kx^{-b}$, where x symbolizes word length as expressed in number of syllables and Y is the average number of phonemes per syllable in words of x syllables length in a given, homogeneous text. K and b are constants that may vary for different

[13] Cf. Lees' critical comments (1959, 285ff.) on an older attempt to find a useful cross-discipline analogy between linguistics and thermodynamics based on the notion of entropy: „It is difficult to see, however, how this formal similarity between measuring the elementary message capacity of a source by partition of symbol probabilities and measuring the unavailability of heat energy by partition of atomic states can be pushed any deeper." (293); „To summarize, then: for good reasons, the communications engineer has been led to characterize the utility of a message source or transmission line in terms of the variety of distinct messages which it permits one to identify (with no consideration of the meaning or understandableness of the messages), and the most convenient expression for this measure involves the logarithm of a probability. For independent reasons, the physicist has been led to characterize the irreversibility of natural energy transformations in terms of a thermodynamic property of systems, the entropy, and he has shown that this property is calculable from the distribution of the particles of the system among available energy states, the expression for entropy then involving the logarithm of a probability. Therefore, the expression for selective information-content and for physical entropy are formally similar; in fact, the very same type of expression, involving the logarithm of a probability, may be used in any number of unrelated problems as a measure of degree of equidistribution." (295). A similar critique of useless formal analogies may be launched against modern complexity theory; as Horgan writes in a well-known popular science article (1995): "Too many simulators also suffer from what Cowan calls the reminiscence syndrome. "They say, 'Look, isn't this reminiscent of a biological or physical phenomenon!' They jump in right away as if it's a decent model for the phenomenon, and usually of course it's just got some accidental features that make it look like something."

[14] The sound change took place not only in Old Russian, but in all Slavic languages, with different results in detail. Basically, two vowels going back to PIE **i* and **u* were eliminated in certain positions and merged with other vowels (in Old Russian, *e* and *o*) in all other positions.

texts. As is obvious, Menzerath's Law dictates a monotonic functional relationship between construct and constituent size. Now, in Old Russian before 1000 AD certain syllable-internal phonotactic restrictions precluded this monotonicity. The authors conclude that Menzerath's Law – in its basic form – was not operative at that time; indeed, curve-fitting leads to negative results for texts that were written before the fall of the yers. The general line of their argument makes use of a *ceteris paribus* reading of the law as criticized above: "It is important to remember right from the outset that Menzerath's law, like any other law, *holds only when the necessary preconditions for it hold*."[15] Talk of 'necessary preconditions' is somehow misplaced here since there is no way to actually *specify* those preconditions except in a circular way: If the alleged law fails in a particular case of application, then it is simply *assumed* that *some* unknown precondition does not hold. The phonotactic facts of Old Russian before the fall of the yers do *not* provide any such unfulfilled precondition. They just give some independent *prima-facie*-indication that makes clear from the outset why we may not even *expect* Menzerath's Law to be applicable here. In saying "Menzerath's Law does not hold here *because* the phonotactics render a monotonic relationship impossible" the *because* must not be understood causally, but *epistemically*, as in "Mr. Lees is not ill *because* I saw him playing with his children on the street just an hour ago."

The assumption that Menzerath's Law was not operative in Old Russian before the fall of the yers has two different possible readings that should be distinguished sharply but get somehow blurred in the authors' presentation, as becomes obvious in their surprising claim (2002, 341) that during the development of Old Russian phonology Menzerath's Law never lost its force, functioning as an attractor that tries to gain control but sometimes, as the reader has to surmise, nevertheless loses its power. The two readings I announced are as follows.

a) *Menzerath's Law, as applied to the ratio of syllable length to word length, holds without exceptions; due to interfering factors, however, it may happen that it does not directly show up in the data.* – This reading has its analogue in physics. Thus, Newton's law of gravitation knows from no exception whatsoever within the framework of classical mechanics, but if we investigate the trajectory of a feather falling in the spring air, the law does not, so to speak, 'shine through' the data because of additional complications such as movement and friction of air. Newton's theory of gravitation is not falsified by such an example, as we can make up, if only in principle, an empirical theory that accounts for the additional factors. Vector addition of the effects of these factors to the effects of Newton's $F = \gamma \frac{m_1 \cdot m_2}{r^2}$ should then yield what is actually observed. Now, the understanding just sketched is implicit in Lehfeldt's and Altmann's talk (2002, 333) of the "possibility of anomalies and boundary conditions that disturb the monotonic direction of the curve." Indeed, a standard strategy in QL to cope with failures of alleged *ceteris paribus* laws is to assume that not all causally relevant factors have been found and accounted for in the deduction of the stochastic regularity postulated. This leads to assuming some ill-defined disturbance factor that is reflected in the derivation by means of one or more additional parameters. As the authors show, modified versions of Menzerath's Law (obtained by adding one or even two parameters to the differential equation that the Law obeys) can indeed be fitted successfully to the Old Russian data in question. However, in contradistinction to our example from physics, no independent theory of the assumed disturbance factors is available, that is, the additional parameters do not receive any empirical interpretation. All that we get, then, is a statement to the effect that adding new parameters

[15] Lehfeldt, Altmann 2002, 331. All following translations from the Russian text of this article into English are mine; in the above quotation, emphasis is mine.

to a formula will improve the results of curve-fitting – a mathematical truism with no immediate linguistic implications.

b) *Menzerath's Law, as applied to the ratio of syllable length to word length, is not operative in certain 'extremal' linguistic situations,* such as the one of Old Russian before 1000 AD. In this reading, talk of attractors seems to be more appropriate since the dynamics of a system might, under certain circumstances, be far removed from the system's attractor(s). In standard definitions of the attractor concept, however, once a system has *attained* its attractor state it will simply remain there forever. The only way, then, to explain why Old Russian before 1000 AD had left the 'Menzerath attractor' is to assume that the Old Russian language system was, during a certain period of time, determined by the effects of *another* attractor. Since attractors are abstract characterizations of the dynamics of a system, this amounts to claiming that Old Russian, at a certain stage of development, *changed* the overall look of its dynamics at least twice, losing the 'Menzerath attractor' before 1000 AD (and being forced to 'wander' to another one) and reenacting it afterwards. Of course, this sort of explanation simply shifts the burden of explanation since what we would need now is (i) a theory of the way the dynamics changes, constructing and deleting attractors in the course of time, and (ii) an explication of the presupposed 'normalcy' or 'default character' of the 'Menzerath attractor' that is implied by the 'law' terminology. Put simply, the first requirement says that *if* we claim to have *explained* the fall of the yers by pointing to the default presence of a certain sort of attractor, *then* we must also be able to explain why the default attractor vanished for a certain period of time in the first place. Requirement (ii) is even more delicate since it points to a stipulated asymmetry in the change of the overall dynamics: Why is loss of the 'Menzerath attractor' the marked alternative vis-à-vis its (re-)establishment that Lehfeldt and Altmann, if implicitly, dub a return to normality?

To sum up: None of the mechanisms one might propose in order to explain temporary 'absence' of the operation of Menzerath's Law can be backed up by anything like an empirical theory – the stipulated attractor remains but a *façon de parler*.

For Old Russian texts after the fall of the yer vowels, Menzerath's Law in its basic form can satisfactorily be fitted to the data. The authors conclude (Lehfeldt, Altmann 2002, 338): "In other words, the fall of the reduced vowels was directed at the elimination of these obstacles [for the law, PM]." Here we see an example of a *post hoc, ergo propter hoc* fallacy, that is, of an illicit causal-final reinterpretation of a merely temporal sequence of events: *After* the yers fell, Menzerath's curve could be fitted again, *therefore*, or so the argument runs, the fall of the yers *caused* or *was directed at* reenacting Menzerath's Law. We have no sound reason to take such a conclusion for granted since it is based on an unwarranted reification of the stipulated reason for the *ceteris paribus* regularity: The claim that 'Menzerath's Law holds *ceteris paribus*' effectively gets rephrased as follows: 'Menzerath's Law is a kind of "telos", a "driving force" that is somehow determined to change the dynamic structure of a language system in the course of time.'

Our result, then, is somewhat negative. No cogent explanation of the language change in question has been offered. Note that even if we had an acceptable scientific theory according to which Old Russian was forced to return to a Menzerath-compatible state, we would still stand in need of an explanation of the yer fall as such because the "goal" of reenabling Menzerath's Law could have been achieved by a host of theoretically possible ways. Indeed, Lehfeldt and Altmann reproach other purported 'explanations' of the yer fall for not providing an answer to their crucial question (2002, 328): "But why did these vowels [that is, these and not others, PM] change at all?" Of course, an answer to this question is part and parcel of the explicitly avowed main objective of the paper under discussion, viz. "to find an explanation for the fall of the reduced vowels" (Lehfeldt, Altmann 2002, 330). The authors claim to actually have found such an explanation (2002, 342), although their formal descriptive apparatus (stating that Menzerath's Law can be fitted to Old Russian data only after the yer fall) can not even in principle give an insight into the actual 'mechanism' that effected the reestablishment of Menzerath's Law – an insight that would, amongst other things, require a complex *phonological* treatment of the Old Russian language.

The ontologizing strategy of positing attractors to explain *ceteris paribus* phenomena seems to me to be a special example of a common argumentative fallacy in QL work: A Poisson/optimization/… process can be used to *model* the phenomenon X, *ergo* there must be such a process, and the process is the searched-for explanans or mechanism; more generally: X shows a certain stochastic regularity R, *ergo* there must be an *underlying causal story* for X that *directly implies* R. Following Lees, we may say that this kind of reasoning instances an "as-if fallacy". Lees, quoting an earlier paper by Quastler, acknowledges that certain statistical properties of music and language are equal to those which could be obtained by stochastic processes – „but it is not claimed that words grow by chance accretion of syllables, or that Mozart's musical line is the result of random collisions" (Lees 1959, 288).[16]

A famous example already discussed in detail by B. Mandelbrot, R. Lees[17], G. Herdan and many others is 'Zipf's Law'. In recent publications Wentian Li has argued anew that "Zipf's law is not a deep law in natural language as one might first have thought" and that "Zipf's law does not share the common ground with other scaling behaviors", emerging instead from ultra-general stochastic premises that hold as well for randomly generated texts (Li 1992).[18] In Mandelbrot's and Li's interpretation, Zipf's Law simply says that natural language texts typically behave, from a stochastic point of view, *as if* they were the output of a random character source. Naturally, this does not mean that such texts *are* such an output. Once again, the search for a mechanism 'behind' the stochastic regularity is determined to fail.[19]

6. Conclusion: Theories and models in the 'system-determined' sciences

Difficulties in quantitative and stochastic modeling similar to those outlined above with respect to QL arise as well in other numerically oriented branches of science. Econometrics

[16] Lees characterizes the epistemic gain from quantitative linguistic models of his time by pointing out that „…the statistical behavior of words in a text, as specified by the explanatory model given, though it results from the operation of ‚known' micro-behavior (i.e. the application of detailed grammatical rules, sociological and psychological determinants of vocabulary, etc.), could also have resulted from the operation of the probability model" (Lees 1959, 287).

[17] Lees summarizes Mandelbrot's famous mathematical discussion on Zipf's Law, pointing out that the law „says merely that whatever the micro-behavior may be that determines our choice of words (what we like to talk about, the grammatical constraints of our language, etc.), it results in an essentially random placement of spaces" (Lees 1959, 287).

[18] Cf. Miller's (in)famous remarks on Zipf's Law in his 'introduction' in Zipf's *Psycho-Biology of Language* (Miller 1968): "Faced with this massive statistical regularity, you have two alternatives. Either you can assume that it reflects some universal property of human mind, or you can assume that it represents some necessary consequence of the laws of probabilities. Zipf chose the synthetic hypothesis and searched for a principle of least effort that would explain the apparent equilibrium between uniformity and diversity in our use of words. Most others who were subsequently attracted to the problems chose the analytic hypothesis and searched for a probabilistic explanation. Now, thirty years later, it seems clear that the others were right. Zipf's curves are merely one way to express a necessary consequence of regarding a message source as a stochastic process."

[19] Lees remarks: „The fact that natural language texts are fair approximations to such random sequences shows merely that linguistic constraints, stringent though they seem to be, still permit sufficient variety in a very long text to approach the ideally random distributions. We see then that the only thing about such frequency distributions which is of immediate interest to the linguist is precisely the departures of natural language texts from the ideal distributions" (Lees 1959, 285).

would seem to be a case in point. In a critique of probabilistic econometric modeling, R.-E. Kalman (1980, 1983) defends a methodological distinction between 'natural' and 'system-determined' sciences. The natural sciences – again with the standard example of theoretical physics – deal with laws of nature in a strict sense that hold regardless of which system is actually considered and that form the backbone of a *theory-driven* approach to empirical phenomena. System-determined sciences such as economics and engineering, on the other hand, venture a *data-driven* approach to highly system-specific regularities. For this reason, their 'laws' often have no validity beyond the specific sort of system to be described and may contain parameters that are neither universal constants nor liable to a general theoretical interpretation. This leads to a remarkable situation in econometrics where mathematically simple formulas with only weak theoretical motivation often turn out to be superior to sophisticated, theory based systems of differential equations when it comes to predictive capacity.

It seems to me that the situation of QL is similar. While the models developed so far do possess statistical significance, theoretical underpinnings remain vague and weak. The *methodic side* of QL research work is close in spirit and in its formal aspects to the rough-and ready inductive generalizations of statistical modeling in the social and economic sciences, whereas its *rhetoric* is that of a super-general, if virtually non-existent, theory of complex, self-organizing systems.

The 'big question' that comes to mind here is whether a "third way" besides a traditional, qualitative understanding of the subject matter of linguistics and the inductive quantitative descriptions of contemporary QL (the empirical side of which rests entirely on qualitative notions in a poorly understood way) is conceivable at all. Recent contributions to the theory of complex systems suggest that qualitative-only and even functional treatments of systems may, in many scientific contexts, be both inevitable and explanatorily fruitful. J. Cohen and I. Stewart (1994) outline a theory of complex phenomena arising or 'emerging' from non-linear causal interaction between two or more systems whose internal dynamics differ so radically from one another that none of the attractors of the individual 'phase spaces' of these systems coincides with any attractor of the combined phase space arising from the interaction. In such cases, the authors argue, the resulting dynamics (which cannot be described from the point of view of the contributing systems) will develop according to simple patterns that are, in a well-defined sense, independent of the complex internal details of each of the involved systems and the specific boundary conditions of the interactions as such. The authors coin the term *complicity* for this kind of interaction. Complicity-driven dynamics can, in many cases, be subjected to a merely qualitative or functional explanation; in other words, the dynamics of the combined system is not reducible to aspects of the dynamics of the systems that form its parts.

To take a favourite example of the authors', the extremely complex evolutionary interaction between the phase space of the internal microbiological and genetic apparatus of the higher living beings on the one hand and the behavioural phase space of the macrophysical interactions of these living beings with each other and with their natural environment has, over and over again, led to the development of *wings* the anatomy and historical morphology of which differ radically from species to species. No look at the intractable details of the evolutionary development of wings in different flying species will give us a deeper *explanation* of the overall fact that flying is reinvented by evolution again and again. Only a coarse-grained, functionally minded explanatory strategy of the "capability of flying enhances overall survival chances" will do here. No reductionism is available.[20] Linguistics, whether 'qualitative' or 'quantitative', possibly faces similar problems.[21]

[20] Cf. Lees (1959, 298): „Reduction of sentences to observational vocabulary and reduction of theories to the vocabulary of physics are usually considered to be independent; indeed, most logical empiricists have by now

Two different ways out of the science-theoretical dilemma of QL and toward a “third way” seem to be conceivable. On the one hand, some of the well-attested stochastic regularities that have been found to date might turn out to be quantitative analogues of the descriptive, non-reductive patterns of complicity as assumed by Cohen and Stewart. In this case, the role of QL laws in future linguistics would be a more mundane, modest one than hitherto assumed; qualitative and quantitative research would simply coexist and be directed at different goals and purposes. QL would not be able to find the deep, hidden mechanisms by which the evolution of linguistic communicative processes proceeds. My exemplary remarks on Zipf’s Law point in this direction.

On the other hand, if a mathematical treatment of the way qualitative linguistic entities such as words, syllables and constituent structures emerge evolutionarily is more than a self-contradictory hope, then it is precisely this mathematics of the ‘complicitary’ *qualitative* concepts we linguists live by that would have to lay the foundations for a mature Quantitative Linguistics.

The preceding remarks are not meant to be an all-or-none deconstruction of the remarkable achievements of QL. Rather, the main thrust of the criticism advanced here consists in noting that the most difficult problems of the discipline are still ahead, waiting to be solved – something most QL adherents will be willing to agree to. The positivist outlook on science that is still fashionable in QL work and the over-estimation of the paradigm of fundamental physics might be an obstacle to solving the most pressing problem, viz. that of bridging the gap between traditional and probabilistic-quantitative modes of thought in linguistics. The favorite quotation of QL, Bunge’s “every thing abides by laws” (1977, 17) is indicative of a fallacy linguists have perhaps fallen prey to just too easily. While Bunge’s dictum is devoid of sense if not taken in a down-to-earth, normative reading (“if you want to do science, try to find regularities wherever you can”), it suggests that simple laws of the sort found in certain natural sciences have to underlie each and every phenomenon of the observable world in such a way that the phenomena in question become mathematically derivable, if perhaps only “in principle”, from a small set of simple equations.[22] To get rid of this admittedly enchanting idea might be a difficult, but necessary step of emancipation from an obsolete paradigm of scientific research.

References

Altmann, G. (1980). Prolegomena to Menzerath’s Law. *Glottometrika 2, 1-10.*

Altmann, G. (1985). Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft 4, 139-155.*

Altmann, G. (1993). Science and Linguistics. In: Köhler, R., Rieger, B.B. (eds.), *Contributions to Quantitative Linguistics: 3-10.* Dordrecht: Kluwer Academic Publishers.

abandoned reductionism in respect to observational terms, while they still often pursue the goal of reduction of sciences to some basic science such as physics. There is, however, a great similarity between the two kinds of reduction, and there is no more reason to believe that the theoretical terms of biology will be eliminable in favor of terms of biochemistry, biophysics, or any other discipline than there is to believe that any theoretical terms in any science are eliminable at all.“

[21] An application of Cohen’s and Stewart’s thoughts to the realm of (meta-)linguistics is hinted at in Meyer 2003. There, the complicity is proposed to arise from the interaction of systems of the following two types: (i) the neurophysiological internal organization of communicating individuals and (ii) their behavioral interaction.

[22] Cf. again Horgan 1995: “Artificial life-and the entire field of complexity-seems to be based on a seductive syllogism: There are simple sets of mathematical rules that when followed by a computer give rise to extremely complicated patterns. The world also contains many extremely complicated patterns. Conclusion: Simple rules underlie many extremely complicated phenomena in the world.”

Altmann, G. (1996). The nature of linguistic units. *Journal of Quantitative Linguistics 3, 1-7.*

Altmann, G. (1999). Von der Fachsprache zum Modell. In: Wiegand, H.E. (ed.), *Sprache und Sprachen in den Wissenschaften der Gegenwart: 294-312*. Berlin: de Gruyter 1999.

Altmann, G. (2002). Zipfian Linguistics. *Glottometrics 3*, *19-26.*

Altmann, G., Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten. *Glottometrika 15*, *166-180.*

Altmann, G., Erat, E., Hřebíček, L. (1996). Word Length Distribution in Turkish Texts. *Glottometrika 15, 195-204.*

Altmann, G., Lehfeldt, W. (2002). Review of: Haspelmath, M. (2000). Optimality and Diachronic Adaptation. *Zeitschrift für Sprachwissenschaft 18.2*, *180-268* [includes discussion articles by several contributors]. *Göttingische Gelehrte Anzeigen 254*, *123-136.*

Altmann, G., **Lehfeldt, W.** (to appear). Review of Peter Siemund (ed.): Methodology in Linguistic Typology. *Sprachtypologie und Universalienforschung 53, 2000, Heft 1.* Berlin: Akademie-Verlag, 123. S. (To appear in *Göttingische Gelehrte Anzeigen*)

Altmann, G., Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Balzer, W. (1997). *Die Wissenschaft und ihre Methoden: Grundsätze der Wissenschaftstheorie. Ein Lehrbuch*. Freiburg (Breisgau): Alber.

Balzer, W., Moulines, C.U., Sneed, J. (1987). *An Architectonic for Science*. Dordrecht: Reidel.

Beekes, R.S.P. (1995). *Comparative Indo-European Linguistics. An Introduction*. Amsterdam, Philadelphia: John Benjamins.

Best, K.-H. (1999). Quantitative Linguistik: Entwicklung, Stand und Perspektive. *Göttinger Beiträge zur Sprachwissenschaft 2*, *7-23*.

Bunge, M. (1977). *Treatise on Basic Philosophy. Vol. 3: Ontology I: The Furniture of the World.* Dordrecht: Reidel.

Bunge, M. (1995). Quality, Quantity, Pseudoquantity, and Measurement in Social Science. *Journal of Quantitative Linguistics 2*, *1-10*.

Chomsky, N. (1978). A Theory of Core Grammar. *Glot 1*, 7-26.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Cohen, J., Stewart, I. (1994). *The Collapse of Chaos. Discovering Simplicity in a Complex World*. New York: Viking.

Earman, J., Roberts, J. (1999). Ceteris Paribus, there is no Problem of Provisos. *Synthese 118*, *439-478.*

Grotjahn, R., Altmann, G. (1993). Modelling the distribution of word length: Some methodological problems. In: Köhler, R., Rieger, B. (eds.), *Contributions to Quantitative Linguistics.* Dordrecht: Kluwer, *141-153*.

Haken, H. (1978). *Synergetics*. 2nd edition. Berlin: Springer.

Hempel, C.G. (1988). Provisos: A Problem Concerning the Inferential Function of Scientific Laws. In: Grünbaum, A., Salmon, W. (eds.), *The Limits of Deductivism*. Berkeley, Ca.: University of California Press. 19-36.

Horgan, J. (1995). From Complexity to Perplexity**.** *Scientific American 272*, *74-79.*

Itkonen, E. (1978). *Grammatical Theory and Metascience*. Amsterdam: Benjamins.

Itkonen, E. (1983). *Causality in Linguistic Theory. A Critical Investigation into the Philosophical and Methodological Foundations of 'Non-Autonomous' Linguistics*. London: Croom Helm.

Kalman, R.-E. (1980). A System-Theoretic Critique of Dynamic Economic Models. *International Journal of Policy Analysis and Information Systems 4(1)*, *3-22*.

Kalman, R.-E. (1983). Identifiability and Modeling in Econometrics. In: *Developments in Statistics*, vol. 4: *97-136*. New York, London: Academic Press.

Kanitscheider, B. (1998). Vorwort: Philosophische Reflexionen über Chaos und Ordnung. In: Peitgen, H.-O., Jürgens, H., Saupe, D. *Chaos. Bausteine der Ordnung: 1-33*. Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Lehfeldt, W., Altmann, G. (2002). Padenie reducirovannykh v svete zakona P. Mencerata [The fall of the reduced vowels in the light of Menzerath's law]. *Russian Linguistics 26(3)*, *327-344*.

Lees, R.B. (1959). Review of: Apostel, L., Mandelbrot, B., Morf, A.: *Logique, langage et théorie de l'information*. Paris. *Language 35(2)*, *271-303*.

Li, W. (1992). Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE Transactions on Information Theory 38(6)*, *1842-1845*.

Meyer, P. (2003). *Gebrauch und Struktur. Zu den pragmatischen Grundlagen grammatischer Beschreibung*. Berlin: Logos Verlag.

Miller, G.A. (1968). Introduction. In: Zipf, G.K. (1968), iii-x.

Mott, P. (1992). Fodor and Ceteris Paribus Laws. *Mind 101*, *333-346*.

Stephan, A. (1999). *Emergenz. Von der Unvorhersagbarkeit zur Selbstorganisation*. Dresden: Dresden University Press.

Suppes, P. (1998). Pragmatism in Physics. In: Weingartner, P., Schurz, G., Dorn, G. (eds.), *The Role of Pragmatism in Contemporary Philosophy.* Wien: Hölder-Pichler-Tempsky.

Weizsäcker, C.F.v. (1985). *Aufbau der Physik*. München: Hanser.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1*, *98-106*.

Wimmer, G., Altmann, G. (1994). The theory of word length: some results and generalizations. *Glottometrika 15, 110-129*.

Wimmer, G., R., Altmann, G. (2003). Unified Derivation of some Linguistic Laws. In: Altmann, G., Köhler, R., Piotrowski, R.G. (eds.). (to appear) *Handbook of Quantitative Linguistic*s. Berlin: de Gruyter.

Zipf, G.K. (1968). *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* Cambridge/Mass.: M.I.T. Press. 2nd printing.

Technology, Ease, and Entropy: A Testimonial to Zipf's Principle of Least Effort

Jeff Robbins[1]

Abstract. Evidence for the truth in George Kingsley Zipf's Principle of Least Effort can be found in the deep attraction we have for anything promising us an easier route. The selling point of virtually all technology is the promise of new means to ease. But, beneath the vast glittering surface in the sea of hype runs a dissipative current. With so many things capitalizing on our aversion to effort, individually and collectively, as a species we're losing it because we're not using it. Seeding that awareness is the first step in reversing the flow.

Keywords: *G.K. Zipf, Zipf's Law, The Principle of Least Effort, effort, technology, dissipative structure, exergy, entropy, the 2nd Law*

The Principle of Least Effort

> "I don't think necessity is the mother of invention - invention, in my opinion, arises directly from idleness, possibly also from laziness. To save oneself trouble."
>
> Agatha Christie

Some years ago, I was reading a slim volume called *Entropy and Art: An Essay on Disorder and Order* by the psychologist, Rudolph Arnheim. In his discussion of "psychic economy," Arnheim casually mentioned a reference to *Human Behavior and the Principle of Least Effort* by the late Harvard linguist/psychologist, George Kingsley Zipf. Intrigued by the title, I set out to find it. The search proved daunting. At one time, the 42nd Street Research Library in New York had the book, but somebody had made off with it. Finally, after a lot of effort, I managed to find a copy in the unlikely stacks of Princeton University's Population Research Library, then located in the basement of what looked like a fraternity house with a cannon out front.

I became a regular at the library, reading a chapter or two each bi-weekly visit from my home on Long Island (a big trip) in a carrel next to one of the few windows. The library was always quiet, frequented, it seemed to me, not very frequently. With too many of us on the planet consuming too much too fast, the library, a major center for population research, should have been bustling. Instead, the librarians were always happy to see a returning customer. The carrel was always free, eagerly waiting just for me, so it seemed.

[1] Address correspondence to: Jeff Robbins, P.O. Box 335, Long Beach, NY 11561 USA.
E-mail: jhrobbins@erols.com

To say that Zipf's treatise was dense is an understatement. It was chock full of case studies traversing an awesome diversity of fields - linguistics, psychology, economics, geography, international relations, you name it - with tables, graphs, and equations galore. In a time before computers, Zipf must have put in an enormous effort to prove to the world that we humans orient our lives, whether we realize it or not, around the prospect of minimizing effort.

Zipf's argument was that we seek out paths to goals that promise the least effort as an energy preserving tactic.[2] Effort burns energy. If you don't know when or where your next meal would arrive, achieving goals such as securing that next meal with the greatest "efficiency," i.e., the least expenditure of energy, could spell the difference between life and death. It made excellent survival sense and is in fact a tactic we share with most, if not all, of our animal cousins. Watch a cat cross the street when she spies no dogs, cars, or humans. She takes her good time. Animals - especially predators - are lazy, and for good reason. They don't waste what's precious.

Zipf based his contention on The Principle of Least Action in physics. By minimizing "action" – the difference between kinetic and potential energy averaged over time – you could predict the behavior of all kinds of interesting things. Substitute effort for action and it's ditto for human behavior. Mind you, it's not that we can't or won't run marathons or climb Mount Everest. The key words, when it comes to decisions about effort, are "average" and "expectation." Just as the big cats burn up the ground securing their meals, then sleep away their bulging bellies, so we too can exert huge efforts in bursts. It's just that with us it's a lot more complicated than sprinting after a zebra. We work like crazy as twenty somethings with the prospect of being able to retire a Silicon Valley millionaire at age thirty and then do what we want (it doesn't always work out).

The Draw of Ease

> "Lest people feel smug about their diligence, evolutionary biologists are discovering that animal inactivity is almost never born of aimless indolence, but instead serves a broad variety of purposes. Some animals sit around to conserve precious calories, others to improve digestion of the calories they have consumed. Some do it to stay cool, others to keep warm. Predators and prey alike are best camouflaged when they are not fidgeting or fussing. Some creatures linger quietly in their territory to guard it, and others stay home to avoid being cannibalized by their neighbors.
>
> "So while there may not be a specific gene for laziness, there is always a good excuse."
>
> Natalie Angier (1991)

The attraction we have for virtually anything that promises to make life easier is so natural, we don't give it a second thought. If you say something "is easier," you need say no more. Easy is good. Easier is better. Easiest is best. If you say it's "simple" or "simpler," if it's "fast" or "faster," if it "saves time" that's great, we want it. We love convenience. We like to be comfortable. If one company produces widgets more efficiently than another, we like that company, we'll buy the stock. If a review says a book, a movie, a show, is exciting, we want to read it, watch it, go to it.

Smack at the beginning of the 1999 beach season, an ad appeared in *The New York Times* for what was then a new book by crime writer Andrew Klavan. Publisher's Weekly declared *Hunting Down Amanda* "Immensely Exciting…Klavan Bows Down To No One For Whiplash

[2] Actually since energy, as the First Law of Thermodynamics tells us, is always conserved, what is being preserved is "exergy," the usable, potent, transformable, part of energy.

Plotting and Page Whirling Suspense…"[3] Why are we drawn to immensely exciting / whiplash plotting / page whirling suspense? And the answer is the promise of an easy read; just something to pass the time while getting a tan, no (mental) assembly required.

But, if Zipf was onto something, why don't we hear more about him and his proposal? Why, isn't something as powerful and universal as the urge to minimize effort universally recognized? Why isn't it taught from grade school on up? With the vast mental and capital resources of the global scientific empire, one would expect to find a well funded glut of Departments of Least Effort with hoards of postdocs, busy as beavers, delving deep into the history, evolution, and consequences of sloth (actually, we may need a new metaphor; according to Ms. Angier, scientists have found the beaver not all that busy after all). Alas, there is no such department anywhere to be found. Maybe a name change to something like Department of Energy Efficiency Studies might work.

Why don't we see what Zipf saw? The answer is that the urge to minimize effort is so deeply ingrained, it's like air; we take it for granted; it fails to enter conscious thought. Subconsciously we choose the path that appears to require less effort if it gets us where we want to go, all other things being equal. We have to go out of our way not to. Case-in-point: In New York City's Penn Station there are two levels. The lower level is for the Long Island Railroad and the upper level for Amtrak and New Jersey Transit trains. To go from the lower to the upper level you either have to climb eighteen steps or take the escalator. One rush-hour morning, it struck me that even though the number of steps wasn't that great, almost all the commuters going from the lower to the upper level took the escalator.

The following week, I came back with a camera and took pictures. What I found was that, though few of them were carrying bags, commuters would rather wait on line - even a long line - to take the escalator rather than climb those eighteen steps. One would think that they would climb up the short flight knowing that getting the heart rate up with a little stair climbing would be good for them. They've been sitting for an hour or more on a train, or will, or are on their way to sit for seven or eight hours at a desk typing, clicking, staring at a VDT. But, once again, one would think wrong. Sure, there was the occasional stair climber. Maybe she was in a rush; maybe she thought about the exercise; maybe she did a quick estimate and decided waiting on line took more effort than climbing the stairs.[4]

Shades of Effort

Infrared film in a camera zeroes in on the heat signatures of animals, people, buildings, cities. By filtering out all but infrared "heat" wavelengths of light, you can see things that would be hard or even impossible to see with ordinary film. An effort filter, fitted over the camera of experience, may reveal signatures and patterns hard or impossible to see in all its full blown complexity. Everybody choosing to take the escalator has a different bag of motives. But, cut out everything but effort and the sheer difference in numbers gravitating to the free ride escalator rather than the effort demanding stairs is telling us something.

[3] *New York Times*, 24 June 1999, E26.

[4] Over a number of months, in half-hour intervals, at different times of the day, I actually counted the numbers taking the escalator vs the stairs. Eliminating all those lugging pullbags, or people with obvious disabilities, my tally so far is 10516 taking the escalator and 1687 climbing the stairs, a ratio of 6.3 to 1.

Shortcuts

Putting on our effort shades, we see that in our daily lives there are countless cases-in-point, little things we pay no attention to because we have too many more important things on our minds to notice. We come to a corner and there's a vacant lot. Heading at a perfect 45 degree angle is a short cut through the grass. Like taking the escalator, cutting diagonally across the lot, makes sense because it is the path of least effort. Unlike the escalator, however, which may take longer than climbing the stairs, it is also the path of least time. It makes sense to cut across the lot unless, of course, you're out for a run.

Do joggers cut diagonally across a lot? From my own experience jogging, I definitely feel the urge to take the shortcut and get done with my run sooner. Why? Because there's always more or less discomfort when you're running and your body wants the pain to end. Something else to notice; just as we like the free rides of escalators - I do - we fail to see that the ride really isn't free. The toll is just hidden from us since we don't explicitly pay for the energy needed to carry us up and down. It's a little thing, but the escalator runs on electrical energy that, in all likelihood, is generated by consuming a fossil fuel that pumps out an infinitesimally added increment of the global warming, greenhouse gas, CO_2. Along the diagonal path that saves us time and effort, the grass is gone, trampled underfoot.

"Rollaboard"

In my investigation of escalator vs stairs in Penn Station, I excluded everyone taking the escalator with a pullbag from the tally because it seemed unfair to expect people to bump them up the stairs unless they absolutely had to. According to a *USA TODAY* cover story by Bruce Horovitz, the luggage industry calls pullbags "rollerboards," a takeoff on the original trademarked name "Rollaboard" coined by its inventor, former Northwest Airlines pilot, now semi-retired multi-millionaire, Robert Plath (Horovitz, 2003). The enormous, industry saving, success story of rollerboard luggage should be evidence enough for the truth in Zipf's proposition.

Back in 1988, when no one knew from rollerboards, Plath felt a compelling need for some kind of compact carry-on bag with wheels to make his pilot's life easier. With nothing on the market at the time, he came up with a stand-up carry-on bag with an extendible handle and wheels and called it the Rollaboard. The name matched perfectly its intended application. It took off at a pace that stunned its inventor. But why? And the answer is it brilliantly appeals to the urge to minimize efforts in virtually all - but not all - respects. In fact, aside from the occasional aisle seat passenger whose elbow, knee or head gets whacked and the delays in takeoff as passenger move around the cabin trying to find an overhead compartment with some space to jam them in, the only significant complaint Horovitz notes about these bags is the amount of effort needed to stuff everything into the carry-on 22 inch version and then live with everything they haven't taken along. "'These carry-on bags,' said one traveller who owns one, but always checks it in, 'have grown in proportion to the laziness of the average traveller.'"

The rollerboard, Rollaboard, pullbag, whatever you want to call it, began racking up sales as soon as travellers spied them because they're clearly easier than what they're replacing. It's easier to roll a bag than carry it either in hand or with a strap over your shoulder. Apart from the risk that your bag will end up in Chicago and you in LA, travellers don't like to wait for checked baggage. Waiting is effort, effort that almost everyone considers a waste of time. Before the rollerboard, travellers checked bags because they had to. Many checked duffel bags that could fit into overhead bins because they didn't want to lug them around the airport. They took a lot of things with them in big suitcases because they knew they were checking them in. They took a lot of things along because it was easier than trying to spend time thinking about what they really did or did not need; easier than having to make do at their destination. But now, with the 22 inch rollerboard that can fit into overhead compartments that airlines, as Horovitz reports, spent $50 million to accommodate, the effort of having to wait for checked luggage coupled to the risk of loss has swamped the added effort of having to think harder about what you can do without and then make do.

There seem to be few drawbacks to effort relieving rollerboards. Some of the most prominent – other people take them when you do check-in the bigger ones because they're all pretty much the same shape and black – are being corrected. Yes, they're noisy and the wheels squeak and break. The only other drawback, which only a crazy exercise fanatic might see as such, is the loss of strength accompanying the disappearance of the need to carry your luggage. Carrying bags, however unpleasantly you look at it, requires muscle and builds muscle. It's a little thing, but think about how much muscle strength has been lost to the nation now that most travellers pull rather than lug? Since waiting is an effort, despite the bad press, it builds a certain kind of muscle. Call it the patience muscle. Add patience to the muscle that's been lost thanks to the Rollaboard.

With almost everything else also relieving us of the need to contract some muscle – even the teeny muscle required to roll down a car window, or actually get out of your car and walk into a restaurant, or get off your butt to change channels, or endure that bit of unpleasant cold on frosty winter mornings when once upon a time you didn't have a remote start - what is happening to us?

Pandering to The Principle / Selling The Goods

> "In 1970, Americans spent about $6 billion on fast food; in 2000, they spent more than $110 billion. Americans now spend more money on fast food than on higher education, personal computers, computer software, or new cars. They spend more on fast food than on movies, books, magazines, newspapers, videos, and recorded music — combined."
>
> *Fast Food Nation* (*Schlosser, 2001:3*)

In the Letters section of same issue of *USA TODAY*, Alison Kretser, M.D. from Washington, D.C., commented on an article claiming that legislators are trying "to limit soft drinks, sugary snacks at schools" in an effort to combat the growing obesity in the nation's children. Although they say they are trying to encourage a more active lifestyle as well as encourage the eating of healthier food, the state bills mentioned in the article are mainly about bans and restrictions on junk food and drink. But these bans and restrictions don't give parents or schools the needed funds and resources to help them nourish better eating and exercise habits. Says Dr. Kretser, "Numerous studies have shown that obesity is related to any number of lifestyle decisions and habits. In fact, research shows that our children's level of activity has dropped dramatically while calorie consumption has remained roughly the same since the late 1960s."[5] She says that pointing fingers or assigning blame to any particular food or beverage is not the answer. A more balanced approach is called for. Towards that end she proposes:

[5] Alison Kretser, M.D., *USA TODAY*, 19 February, 2003, Letters, 11A.

- "Providing sound information on nutrition to parents, students and teachers.
- "Encouraging more physical education and recreational opportunities.
- "Funding the research we need to determine ways to encourage healthy lifestyle choices."[6]

This is all very good advice and Dr. Kretser is not at all alone in her recommendations. But consider what she is proposing in the context of everything in our individual and social environments that panders endlessly and successfully to minimizing or eliminating the kinds of efforts implied in her proposals. Reasonable and balanced as they seem, in reality they go entirely against the steamroller of consumption that keeps a really big chunk of our economy rolling.

Sound information on nutrition; eat healthy; great. But what about the $110 billion fast food industry, an industry that sells itself, hook, line, and sinker, on eliminating both sides of the effort equation. They make it as easy as possible to get food. They make it as easy as possible to eat food by engineering and standardizing the taste so that it hits the taste buds hard and consistently. The fast food industry, despite what it may say to kill any lawsuits claiming fast food is making kids obese, does not want its customers to think about what they are consuming (if they did, there's a good chance they wouldn't be choosing fast food).

Get exercise. Expand physical education and recreational opportunities. Sounds wonderful and it is. But now look at the massive industries that depend on kids staying home, playing video games, surfing the internet, and above all else watching television so that they can be sure to catch the ads on sugar loaded drinks and cereals and, let's not forget, fast food.

The rollerboard is smart, it makes perfect sense, it’s much easier than what it replaces. But so is everything else.

Make it Easy, Make it Simple, and Make it Fast

> "Curbside Takeaway / No Rules. Just Right to your Car
> "Call up. Pull up. We'll bring the food straight to your car."
>
> Ad for *Outback Steakhouse*

The attraction of ease is as universal as those who want to capitalize on it. At least one of the reasons why nobody seems to be talking about the powerful attraction we have to anything promising less work is that the awareness is not wanted. Gaining an edge in sales too often hinges on making it seem that your own product is easier than your competitors. If consumers were to start thinking about just how much the promise of easy is exploited; if they began mulling over what too much ease, too much indulged, might be doing to them, to their kids, to their community, to other species, or to the entire planet, it might get harder to sell them stuff.

One of the easiest - there you go - ways to get an idea of just how automatic the expected response to simplicity, ease, and speed, is to do an online *Google* search on the words, "Made Simple". Here's a sampling from a search done in December 2002 that came up with 5,500,000 sites matching "Made Simple".

> Simple Machines Made Simpler. Hematopathologic Phenotypes Made Mockingly Simple. Success Made Simple. Bankruptcy Made Simple. Credit Repair Made Simple. Extraordinary Meetings Made Exceedingly Simple. Boat Cosmetics Made Simple. Wildebeest Migration- 'The Migration Made Simple.' Filing Made

[6] Ibid.

> Simple. Mall Shopping Made Simple. Marxism Made Simple. The Great Outdoors Made Simple. Tragedy Made Simple. Celsius and Fahrenheit Conversions Made Simple. Privacy Made Simple For High-Tech Minds. Black Holes Made Simple. Complex Made Simple…

Search on "Made Easy" and *Google* responds with 5,970,000 sites. A sampling:

> Embryos Made Easy. Hernias Made Easy. Southern Living Microwave Cooking Made Easy. Orthopedics for Poultry Made Easy for Beginners. Multiplication Made Easy. Attention Deficit Disorders Made Easy. Tarantulas Made Easy. Arabic made Easy. Computers Made Easy For Senior Citizens. Parking Fines Made Easy. Meditation Made Easy. Hacking Hotmail Made Easy. Feng Shui Made Easy to Improve Your Luck! Work Out Planning Made Easy. Transvestite Transformation - Make-Up Made Easy. Doing Good Made Easy. Move-In Made Easy. Complex Numbers Made Easy. The Rules Made Easy AKA Rules For Idiots Video. Object Encapsulation Made Easy. Job Search Made Easy. Baltimore Made Easy. Shopping Made Easy. Weddings Made Easy, Computer Software. Chrysanthemum Breeding Made Easy. Visual Basic.Net Tracing, Logging and Threading Made Easy. Finding a Niche Made Easy. Bangs Made Easy. Men Made Easy E book. Opium Made Easy. Muscles Made Easy. Menopause Made Easy. Life Made Easy.

Not only do we want things to be simple and easy we want them to be fast. What was it the ad said? Ah, yes, "We hate to wait." Put "fast" in the *Google* box and the engine responds in 0.05 seconds - a response time that is truly amazing - with "about" 35,200,000 sites all of which have "fast" somewhere embedded. "Welcome to Fast Search and Transfer" says the header of one site. "FAST products, FAST Technology, FAST Success Stories." The dust jacket on James Gleich's 1999 book, *Faster: The Acceleration of Just About Everything*, says: "If one quality defines our modern, technocratic age, it is acceleration. We are making haste. Our computers, our movies, our sex lives, our prayers – they all run faster now than ever before." In ever greater numbers we have become the Type A man and woman in a rush to do more and more in less and less time. "We have become a quick-reflexed, multitasking, channel-flipping, fast-forwarding species." In ever increasing numbers we push the DOOR CLOSE elevator button.

It takes less effort to wait less than to wait longer. Having to deal with complexity is uncomfortable. Having to do something that's hard means having to endure feelings that you'd rather be over. Waiting is pain. Speed is the relief.

"That's Entertainment"

> "For all practical purposes, the U.S. today is a 24-hour, TV entertainment society. Everything in contemporary America is an entertainment, from sporting event to big business, politics, certainly religion, and even academia. If it isn't fun, cute, or packaged in a ten-second sound bite, then forget it. If it can't be presented with a smiling, cheerful, sexy face, then it ain't worth attending to. We're all spectators in a grand entertainment society looking up at the few superstars on the stage who not only perform but stand out enough to be labeled heros of our age. In critic Richard Schickel's biting observation, in contemporary America one is either a celebrity or one is nothing."
>
> *The Unreality Industry (Mitroff and Bennis, 1989: 8-9)*

Like the word “easy,” when someone says “entertainment” we like it, no qualifiers required. Generally we have a limited set of things in mind when we think of entertainment. Movies, Broadway, On and Off, circus, best sellers, magazines, tabloids, TV sex, TV violence, reality TV, sports on TV, wrestling on TV, car chases on TV, video games, pop music, music videos, concerts. We don’t consider escalators, shortcuts, throw-away shavers, one-use cameras, the stock market, state lotteries, guns, bombs, automobiles, air conditioners, motorized lawn mowers, tractors, giant cranes, agriculture, assembly lines, genetically engineered potatoes, fast food, sugar loaded colas, cigarettes, prescription drugs, ski lifts, dictionaries, *Google*, restaurants, packaged steaks, airplanes, trains, boats, buses, remote controls, software, microwave ovens, manufacturing robots, motorized golf carts, schools, calculators, textbooks, chauffeurs, butlers, maids, waiters, chefs, health clubs, nail salons, personal trainers, free weights, treadmills, jewelry, mink coats, actors, celebrities, coaches, teachers, librarians, doctors, lawyers, politicians, and so on, “entertainment.”

But put on those effort shades and you begin to see that, without exception, these seemingly unconnected or marginally connected entities share one thing in common; they all reduce, minimize, or completely eliminate effort. It's true that many of these "entertainers" make some things easier only to make other things harder. Free weights make building muscle easier by offering resistance. From tabloids to texts on Theoretical Physics, the spectrum of demand varies from one extreme to the other. But regardless of where they fall on the scale of effort, all of them succeed or fail on the *promise* of something or someone doing some, most, or all of the work for us. And that's entertainment.

Consider this: What could air conditioning and television possibly share in common? Answer: both eliminate effort.

Air conditioning eliminates the effort, and it is an effort, needed to deal with heat with our own substance. For the millions of years, humans and proto-humans have been around, somehow we managed to live with heat and humidity. Even today, outside of the developed industrialized nations, people still have to go with the flow of hot and humid. They adapt; take siestas; build their homes with lots of free flowing air under big shade trees. But, now, especially in America, arguably the most air conditioning dependent nation in the world, technology does the work of cooling the environment for us. Air conditioning makes living with hot and sticky easy. Just turn the AC up.

Television eliminates the need to exert effort period. Is there anything easier than plopping on your couch and surfing with your remote to see what will most amuse you for the moment? As Robert Kubey and Mihalyi Csikszentmihalyi (1990: 81) reported in their book, *Television and the Quality of Life*, there is virtually no other activity that requires lower levels of concentration, challenge, and skill than watching television. Reading? It takes more work. Thinking? Not even in the same ballpark. Writing? Much more work. The secret of television's overwhelming success is its ability to exploit the Principle of Least Effort. But, the answer is yes, there is something easier than channel surfing for what's on: It’s technology that does the surfing and the recording for you, because it knows what you like, so that you can watch what you want when you want and skip the ads as a plus.

Then again, what could a tabloid possibly have in common with a text on Theoretical Physics? Tabloids exploit everything that hooks eyeballs; celebrity scandals, extraterrestrial abductions, angels to watch over you, eat anything you want and lose 40 pounds in thirty minutes diets, 536 ways to have great sex, you name it. With a hard physics text, it's just the opposite. You have to make a major effort just to fill in the blanks between equations. The only thing you can count on when you see "it follows that," or "a short calculation reveals" is that it does not [immediately] follow that and the calculation will not be short. Nonetheless, the mystery of how one step leads to the next makes it easier for the student to make the effort to get a leg up on theoretical physics. Like free weights, a good teacher motivates her students to put in the effort and in so doing makes that effort easier.

The Web

> '"Tell me about the role of Elizabethan theatre in English literature. Why is it called "Elizabethan"? Please send me all the information right away, as the paper is due tomorrow."'
>
> Student query to Usenet group (Sweetland, 2001)

The Web is a remarkable outcome of technology. It allows us to access information, "to find out about," to connect to a vast, ceaselessly changing reservoir of links as never before in history. Powerful search engines like *Google* can sift through billions of Web sites in a fraction of a second. With so much power at our fingertips, entirely new possibilities for research have come online. But mount the effort filter and what you begin to see is that the foundation on which the entire Web edifice is built is none other than the profoundly simple desire to minimize effort.

As with all technology worth its salt, the Web gives us new freedoms and choices previously unavailable. We can choose to do a Web search or engage in traditional offline research or both. Now we have the choice. Before we did not. This is a plus. A big one. We can now devote our efforts to new and deeper possibilities. The choice is ours. But is it?

For some the answer is absolutely yes. We can take the steps if we want to. All it takes is an idea; "a little exercise might do me good." We can choose the convenience and hard hitting taste of supersized fast food and sugar drinks. But, we can choose healthier fare in healthier amounts. It's up to us. We can let our powerful graphing calculator do most, if not all, of the calculations and graphing and devote our energies to the higher math it still cannot do. The technology gives us the choice and it's up to us to make good use. But is it?

In a recent (least effort) *Google* search on the key words "The Principle of Least Effort," I came across an interesting *Library Link*. Written by James H. Sweetland, its title is "The Need for Guides, Coaches, and Teachers in the Self-Service Information Environment." Speaking to his fellow librarians, what Sweetland is lamenting is an unintended consequence of their exertions to make "ever more 'user friendly systems'." He says that the end result of the ease of doing online research is that students are taking the path of least effort to an extreme and the consequences are spilling over adversely in the quality and depth of their work.

Says Sweetland, "[There is] a growing body of research that most students use extremely simplistic search strategies in electronic sources, notably the typing of a string of the most obvious terms into a search engine." And if the first strategy doesn't work, they just give up without seeking any help. Complaints from Usenet groups and Web sites seem to point to student queries being nothing more than querying the site. The intended positive use of the technology is not being realized because it turns into a substitute for personal effort rather than a liberator allowing effort to be devoted more profitably.

Across the Board is What's Wrong

Effort is what puts order in our minds and bodies. Effort is what puts order in the world. Effort is what counters the universal tendency of things to fall apart, for energy to spread out, to flow downhill like a meandering river from useful to useless. But exerting effort consumes food energy. In a world of scarcity, efficient use of energy is no luxury. So it is that we have this deep seated urge, a survival instinct, that is delighted when something, anything promises us a more energy efficient route to whatever it is we want to accomplish. We're universally drawn to the promise of minimizing the work involved. Always there to remind us that we're burning up energy when we apply effort is pain. If it were not so, we would be cavalier about effort, consuming food energy without knowing it. Since we don't like pain, when a piece of technology comes along that alleviates effort we equate that with the relief of pain and that's good. The escalator saves us the effort of walking up steps. It is more pleasant to get that seemingly free ride, however short, than to endure the discomfort, however brief and mild, of climbing steps. For every person who, for whatever reason, chooses to climb up 18 steps, between six and seven ride the escalator. The escalator is successful. Its creators may not have heard of George Kingsley Zipf, but they would understand very well what he was driving at. They understood The Principle of Least Effort and acted on it.

But smart, efficient, and profitable as it seems to provide products and services that make life easier, there is a problem. The problem is not one or two or ten or even a hundred things making life easier. It is across-the-board everything.

Zipf: A Different View

The name of George Kingsley Zipf, if it is known at all, is linked to the remarkably diverse collection of phenomena gathered under the umbrella of Zipf's Law.[7] While few doubt the validity of Zipf's Law - especially in its more refined incarnations - when experimental results can be gathered and ranked, the connection of the law to the Principle of Least Effort remains, to this day, smoking with controversy. Rather than add more coals to the fire, my purpose here has been to add weight to Zipf's contention that people, individually and collectively, strive to minimize effort[8] by following a different route; one that connects the Principle of Least Effort to our growing dependency on the power of technology.

It bears repeating that we fail to notice the linking of virtually all technology, in its design, implementation, marketing and use, to the urge to find an easier route because it's just so obvious that it's below the level where conscious thought can give it a look. Unfortunately, we can no longer afford not to look.

Whereas the instinct to latch onto anything promising us a more efficient solution once - and for millions of years - made excellent survival sense, today it is an instinct that is out-of-touch with our technology transformed world. Fat is tasty because, in a natural world of scarcity, stored fat could be a life saver. The good taste reminds us of fat's survival value. Today, when food is as hard to get as your local drive-up window, or a click of your mouse, it still tastes good. Once more: tasty fat is easy to eat. Tasty food is easy to get. In America, at least, never has it been harder to burn up the calories in what's easy to eat and easy to get.

When everything conceivable sells itself on the promise of eliminating some kind of effort, and effort is what we need to create and maintain order in our bodies and minds, entropy, the measure of energy's powerlessness, fills the vacuum.

As technology systematically encroaches on the territory of human effort, rendering us more and more dependent on its power to keep us going, in effect, what is happening is that order, and

[7] In its most simple formulation, "Zipf's Law is a relation between the frequency of occurrence of an event and its rank when the events are ranked with respect to the frequency of occurrence (the most frequent one first)." Frequency multiplied by rank equals a constant (Rousseau, 2002).

[8] Zipf defined effort as the "*least average rate of probable work*," (Zipf, 1949: 6)

the power that order confers, is being transferred out of people and human social structures and into the order of technics. When we, the consumer and user, buy into technology precisely because of its promise to make life easier, as the power in the technology goes up the power in the people goes down. Of course, this is the opposite of we're all being told.

The way it's supposed to work is that we're in a win-win situation: as the power in what we use rises ever upward on the stepping stones of latest versions minus one, ipso facto the power in us goes up. But just because we double click a mouse and transform the world, doesn't mean that the power of transformation lies somewhere within us. The power is in the technology. All we did was double click a mouse. The transfer of power from the technics to us is an illusion, one that only becomes apparent when, for whatever reason, the power, like a lost mobile phone, is gone.

Is Technology a Dissipative Structure Feeding on the Draw of Sloth?

> "…even when a person is comparatively at rest, there is still a continual movement of matter-energy into his system, through his system, and out of his system…"
>
> G.K. Zipf (1949: 1)

The 2nd Law of Thermodynamics equates the passage of time with the increase in universal entropy. Overall, entropy never ever goes down. But this ironclad dictum does not prevent entropy from going down locally. If it did we would not be here. Life would never have come about. Reducing entropy in a particular local system (a bounded collection of matter and energy in flux like a cell or a person) means that the energy in that system has greater potential for useful transformation. It has more exergy, the measure of energy's power, much like a weight that has been lifted in a gravitational field and can do useful work when it falls back down.

Even simple physical and chemical systems, open to the flow of matter and energy, manifest the tendency to lower entropy locally by spontaneously self-organizing. A case in point can be found in the appearance of Bénard convection cells in a thin layer of liquid contained between two plates being heated from below. Cell formation represents the spectacular, instantaneous, coordination of in excess of 10^{22} molecules when a critical threshold of temperature difference between the plates is reached. Another example is the BZ (Belosov-Zhabotinski) reaction producing an oscillating chemical clock - yellow, colorless, yellow, colorless... (Nicholas and Prigogine, 1989: 6-28). Spontaneous self-organization occurs in systems that are being maintained far from equilibrium as is the case when the critical temperature threshold is reached and Bénard cells form. The Nobel Laureate physical chemist, Ilya Prigogine, called these self-organizing systems "dissipative structures."

A dissipative structure concentrates order (lowers entropy) in itself by shipping off entropy into its environment. The astonishing self-organization of the Bénard cell can only take place because its high order permits a radical increase in the heat flowing between the plates. The stepped up heat flow is produced by burning fuel at a much higher rate. The escalated pace of burning fuel bumps up the rate of universal entropy production and this is what allows entropy to go down locally. Even though entropy goes down in the Bénard cell, when you take into account *both* the high order system and its environment, overall entropy goes up. The 2nd Law is happy. Bénard cells form.

In an essay "Order from disorder: the thermodynamics of complexity in biology," that appeared in the collection, *What is Life? The Next Fifty Years*, Eric D. Schneider and James J. Kay made a fascinating proposal as to why the 2nd Law allows order to concentrate locally in dissipative structures.[9] The 2nd Law tells us that the universe wants to maximize entropy production in any way it can. If a system, far from equilibrium, can escalate gradient destruction (e.g., hot and cold going to lukewarm) and thus increase entropy production in a major way through the self-organized production of chaos, the 2nd Law will love this and will only be too willing to accommodate. What this is saying is that the evolution of life to ever more complex organisms, biological and social, took place not for the sake of creating or preserving more order in the world, but the opposite. Life both came about and evolved to bring about a more ordered means to unleash chaos; to allow entropy to increase far more than it would have had self-organization never occurred. In the words of Schneider and Kay:

> “[The] emergence of organized behaviour, the essence of life, is now understood to be expected by thermodynamics. As more high quality energy is pumped into an ecosystem, more organization emerges to dissipate the energy. Thus we have *order* emerging from *disorder* in the service of causing even more disorder.” (Schneider and Kay, 1995: 170).

If we extrapolate from Schneider and Kay’s proposition, evolving technology represents the human unleashed mechanism whereby order can be focused as a means of more efficiently satisfying human needs and solving human problems. Technology carries on where biological evolution leaves off. It is a mechanism for the capturing of higher and higher levels of order. But, contrary to the conventional wisdom, the 2nd Law allows technology, loves technology, not for its vaunted capacity to inject order into the human world, but for its ability to escalate chaos.

When it comes to military technology, the chaos unleashing factor in advancing state-of-the-art is not hard to see (or hear). On that first night of "shock and awe," exploding Tomahawk cruise missiles launched from warships far removed from their intended targets lit up the sky over Baghdad. The high and evolving order (low entropy) in military technology is a means for the high and evolving release of entropy as targeted destruction, death, and suffering; in a word, chaos. When it comes to war, entropy is the name of the game.

But then, at least military technology is honest. It might be hard to find someone who would argue that guns, bombs, or missiles are meant to enhance the lives of those they're aimed at.[10] Consumer technology is different. One way or another, it is sold expressly on the promise of enhancing our lives. But is consumer technology really so different than its military cousin?

[9] The idea of creating 'order from disorder' was proposed by the late, great, physicist, Erwin Schrödinger, in a series of lectures given in Trinity College, Dublin in 1943. The lectures were captured in what became one of the most influential "little" books in the history of science. It was called *What is Life?* As Schrödinger saw it, "the problem faced by organisms was how to retain their highly improbable ordered structure in the face of the second law of thermodynamics. Schrödinger pointed out that organisms retain order within themselves by creating disorder in their environment" (Murphy and O'Neill, 1995: 2). See also Schneider and Kay's essay as chapter 12, 161-173.

[10] When the cover story of a best selling supermarket tabloid proclaims that a Titanic baby has been found alive, floating in the Atlantic on an inner tube without food or water, and, mother of all wonders, still a baby since 1912, you never know what people will believe.

Case-in-Point: Television

> "'I think what is probably the biggest sin of the medium as it exists is that so little sticks to your ribs, that so much effort and technology goes into—what? It's like human elimination. It's just waste.'"
>
> Grant Tinker[11]

A number of years ago, I presented a paper with a model of the television producing system as a dissipative structure with its vast consuming public as the environment into which it exported entropy (Robbins, 1989). One of the most significant ways the TV producing system gains power is by dissipating its human environment. How does it do this? It does this by becoming what is arguably the most successful technology in history to capitalize on - you got it - G. K. Zipf's Principle of Least Effort.

For argument's sake, if there are 250 million TV viewers in the USA and each, conservatively speaking, watches television 3 hours per day,[12] the nation spends more than 273 billion hours a year in an "activity" that, as Kubey and Csikszentmihalyi noted, is without peer in the department of lowest demands for concentration, challenge, and skill. Because virtually anything else people do requires more effort than watching television, by eliminating 273 billion hours of order producing effort, in effect television dumps 273 billion hours of disorder, a.k.a. entropy, into the minds and bodies of Americans of all ages. And that's just the tip of the iceberg. Television is not only number one in the elimination of human effort, it is number one as the driver of consumption. Since there has never been a product or service that doesn't come with a 2nd Law toll, just as the self-organizing of Bénard cells radically increases heat flow and therefore entropy production, the TV industry unloads massive amounts of entropy into the biosphere by radically increasing consumption. If we're talking sustainable future, this is not the way to go.

Some Concluding Thoughts

We humans may be dissipative structures, like every other living organism, but we are not like the simple (relatively speaking) Bénard cell whose order self-organizes to increase heat flow and thus entropy production. We can and do produce chaos, but we can also, thanks to effort, impart order to the world and to ourselves. We are creative and destructive. It is the same with technology. It not only amplifies destruction, it enables creation. It really does let us do things we could never do before.

We may be the environment of high tech, but we are not rivers, not the atmosphere, not the oceans, not the soil that have no choice but to accept whatever is dumped. If technology unloads its entropy into us by capitalizing on our weakness for all things promising us less or no effort, we can do something about that. We can inject effort into the equation. We can choose when and how and how much we will deploy the power.

How can we do this? For starters, we can do this by recognizing what Zipf had been trying to tell us all along. By knowing we have this instinctive attraction to anything promising us

[11] Kubey and Csikszentmihalyi (quoting Gitlin (1983) *Inside prime time*. New York: Pantheon, 16). Grant Tinker is a TV mogul perhaps best known as the producer of Mary Tyler Moore, Cheers, and Bob Newhart.

[12] It's difficult to find consistency in estimates of average viewing time. A. C. Nielsen is the rating service most frequently cited, but TV producers, whose advertising rates depend on who and how many are watching, have complained in the past that Nielsen's ratings were too low.

an easier route we can choose to *not* take the escalator. We can choose to press the OFF button. We do not have to push DOOR CLOSE.

Although this work hardly scratches the surface, if the idea of bringing to conscious thought Zipf’s Principle of Least Effort allows us to better tap into the order side of our conversation with technology, the added effort will make good use of the extra energy consumed.

Some rules of thumb. The 2nd Law never ever permits universal entropy to shrink. It allows local reduction only because the self-organized production of chaos generates more of what it wants. Technology, as it continues its exponential evolution, represents nodes of ever increasing power. To satisfy the 2nd Law, it must export entropy into its environment. Military technology does this clearly and honestly. Consumer technology does it subtly by capitalizing on Zipf's Principle of Least Effort. It's not that making life easier is the problem, it's the vast galaxy of things eliminating effort. The systematic global elimination of human effort is the problem. There's a message in the accelerating order being poured into robotics, nanotechnology, and genetic engineering, and it is worrisome. Injecting effort back into our conversation with technology may be crucial if we are to have a future. Awareness is a first step on that road.

References

Altmann, Gabriel (2002). Zipfian Linguistics. *Glottometrics 3, 2002, 19-26.*

Angier, Natalie (1991). Busy as a Bee? Then Who's Doing the Work. *New York Times*, 30 July 1991, C1.

Arnheim, Rudolph (1971). *Entropy and art. An essay on disorder and order*. Berkeley: University of California Press.

Atkins, P.W. (1984). *The second law*. New York: Scientific American Library. Perhaps the best non-mathematical book on how the 2nd Law works.

Bowman, John (2002). Geographic Profiling. CBC News Online, October 16, 2002. http://cbc.ca/news/features/geographic_profiling.html. Interesting application of the Principle of Least Effort in locating fugitives.

Gleick, James (1999). *Faster. The acceleration of just about everything.* New York: Pantheon Books.

Horovitz, Bruce (2003). Taking the 'lug' out of luggage changed everything: Those little black bags on wheels have made a huge difference - in packing and even in jets and airports themselves. *Usa Today*, 19 February, 2003, 2B.

Kauffman, Stewart (1995). *At home in the universe. The search for the Laws of self-organization and complexity*. Oxford: Oxford University Press.

Kubey, Robert and Csikszentmihalyi, Mihaly (1990). *Television and the quality of life. How viewing shapes everyday experience.* Hillsdale, N.J., Erlbaum.

Li, Wentian (1999). Zipf’s Law. http://linkage.rockefeller.edu/wli/zipf/. Li maintains a website of the vast range of phenomena linking to Zipf's Law.

Mander, Jerry (1992). *In the absence of the sacred. The failure of technology and the survival of the indian nations*. San Francisco: Sierra Club. Especially chapter 6: first hand accounts of the impact television on the 20,000 year old Dene nation, Northwest Territories.

Manrubia, Susanna C., Derrida, Bernard and Zanette, Damián (2003). Genealogy in the Era of Genomics. *American Scientist*, 91(2):158-165. Recent illustration of Zipf's Law (as Lotka's Law) in a plot of surname family size versus their frequency in a population.

McKibben, Bill (1992). *The age of missing information*; New York: Random House. Television, whose hallmark is connection, in reality accomplished the opposite.

Miller, George A. (1968). Introduction. In: Zipf, G. K.: *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, Mass: The M.I.T. Press. 2nd printing, iii-x.

Mitroff, Ian and Bennis, Warren. (1989). *The unreality industry. The deliberate manufacturing of falsehood and what it is doing to our lives*. New York, Birch Lane.

Nicholas, Grégoire and Prigogine, Ilya (1989). *Exploring complexity*. New York: WH Freeman.

Postman, Neal (1992). *Technopoly. The surrender of culture to technology*. New York: Knopf.

Prün, Claudia (2002). G.K. Zipf's Conception of Language as an Early Prototype of Synergetic Linguistics. *Journal of Quantitative Linguistics* 6:78-84.

Prün, Claudia and Zipf, Robert (2002). Biographical notes on G. K. Zipf. *Glottometrics 3, 2002, 1-10.*

Reingold, Howard (1999). Look Who's Talking: The Amish are famous for shunning technology. But their secret love affair with the cell phone is causing an uproar." *Wired a*rchive 7.01 (January 1999). Found at http://www.wired.com/wired/archive/7.01/amish_pr_html.

Robbins, Jeff (2000). A Case for the Minimum: Least Effort as Meta-Umbrella International Society for the Systems Sciences, 44th Annual Meeting. Toronto, July 16-22, 2000. ISBN 09664183-5-2.

Robbins, Jeff (1999). Entertainment and the Sustainable Future. Second Interdisciplinary Conference on The Evolution of World Order, http://www.ryerson.ca/~woc/prev_events/conf99/robina.html

Robbins, Jeff (1989). To School for an Image of Television. California State University, Los Angeles: *A Delicate Balance: Technics, Culture, and Consequences* Proceedings, 1989, 285-291. Library of Congress Number 90-84061. IEEE Catalog Number 89CH2931-4.

Rousseau, Ronald (2002). George Kingsley Zipf: life, ideas, his law and informetrics. *Glottometrics 3, 2002, 11-18.*

Schlosser, Eric (2001). *Fast food nation. The dark side of the all-american meal*. Boston: Houghton Mifflin.

Schneider, Eric D. and Kay, James J. (1995). Order from disorder: the thermodynamics of complexity in biology. In: *What is life? The next fifty years*, edited by Michael P. Murphy and Luke S.J. O'Neill. Cambridge: Cambridge University Press.

Schrödinger, Erwin (1944). *What is life?* Cambridge: Cambridge University Press.

Sweetland, James H. (2001). The Need for Guides, Coaches, and Teachers in the Self-Service Information Environment. http://www.emeraldinsight.com/vl=1/cl=3/nw=1/rpsv/librarylink/collection/viewpoints/0101.htm

Tenner, Edward (1996). *Why things bite back. Technology and the revenge of unintended consequences*. New York: Knopf.

Vogel, Jennifer (1997). *Crapped out. How gambling ruins the economy and destroys lives.* Monroe, Maine: Common Courage Press. The explosion of gambling across the USA from casinos to lotteries is powerful evidence for the truth in Zipf's contention.

Vanderburg, Willem H. (2000). *The labyrinth of technology*. Toronto: University of Toronto Press.

Weltfish, Gene (1965). *The lost universe. The way of life of the pawnee.* New York: Basic Books. What have we lost in our headlong rush into the future?

Winn, Marie (1977, 1985). *The plug-in drug.* New York: Viking Penguin. An account of some of the more serious downsides to our addiction to television, especially on children.

Winner, Langdon (1977). *Autonomous technology. Technics-out-of-control as a theme in political thought*. Cambridge, MA.: The MIT Press.

Zipf, George Kingsley (1941) *National unity and disunity. The nation as a bio-social organism.* Bloomington, Ind: The Principia Press.

Zipf, G. K. and Whitehorn, John C. (1943). Schizophrenic Language. *Archives of Neurology and Psychiatry*, June 1943, Vol. 49, 831-851 (Harvard Archives HUG(B)-Z462.72).

Zipf, G. K. (1949)**.** *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley, 1949. Haffner reprint, 1972.

The Gods Must Be Crazy, a film by Jamie Uys on the Kalahari Bushmen, 1980.

Koyaanisqatsi (Life out of balance / A state of life that calls for another way of living), a film by Godfrey Reggio, 1983

Glottometrics 5, 2002, 97-107 *To honor G.K. Zipf*

Oscillation in the frequency-length relationship

Peter Grzybek, Graz[1]
Gabriel Altmann, Lüdenscheid

Abstract. The analysis shows that there is no intrinsic oscillation in the relation between frequency and length of words. The rise of oscillation is caused by using moving averages for smoothing the extremely dispersed data.

Keywords: Frequency-length relation, oscillation

The relationship between the frequency of a word and its length has repeatedly been the object of linguistics studies since Zipf's (1932, 1935) corresponding statements. Subsequent to his hypothesis, stating that the length of a word stands in an inverse relationship to its frequency, many studies have analyzed this problem, based either on texts (or parts of texts), frequency lists, or corpus analyses.

Different models have been suggested to formally describe this particular relationship. In his comprehensive study on the German LIMAS corpus, composed of ca. 500 texts (or parts of texts, respectively), and comprising about one million words, Köhler (1986) tested if the so-called power law, implying the relationship $y = ax^{-b}$, is apt to adequately describe the dependency of word length on word frequency.

As a result of his statistical analysis, Köhler (1986: 137) concluded that his initial hypothesis must not be rejected; in a follow-up study by Zörnig, Köhler, and Brinkmöller (1990: 25), the authors repeat this interpretation, speaking of "a highly positive result". Their interpretation was based on an analysis of variance, yielding $F_{1,158} = 105$ ($p < 0.001$), ($F_{0.01} = 6.8$). Meanwhile, however, it is a well-known fact, that as the sample size increases, the F-test is problematic to reliably interpret results achieved by it. Therefore, it seems to be more reasonable to (at least additionally) evaluate the goodness of regression models by reference to the determination coefficient R^2, although the latter, too, of course, is not free of deficiencies (cf. Grotjahn 1992).

In fact, a re-analysis of Köhler's data shows that his initial fit is far from being good, which is corroborated by the determination coefficient of $R^2 = 0.24$. This poor fit is clearly illustrated in Fig. 1, taken from Köhler (1986: 138).

As a matter of fact, Köhler and his co-authors noticed that, irrespective of the positive F-value, the fit of the exponential function was far from being satisfying. Therefore Köhler (1986: 137) concluded that, given the deviations of the empirical data from the theoretical curve are random, it should be possible to arrive at better results, if one smoothes the data by way of moving averages.

[1] Address correspondence to: Peter Grzybek, Institut für Slawistik, Universität Graz, Merangasse. 70, A-8010 Graz. E-mail: grzybek@uni-graz.at

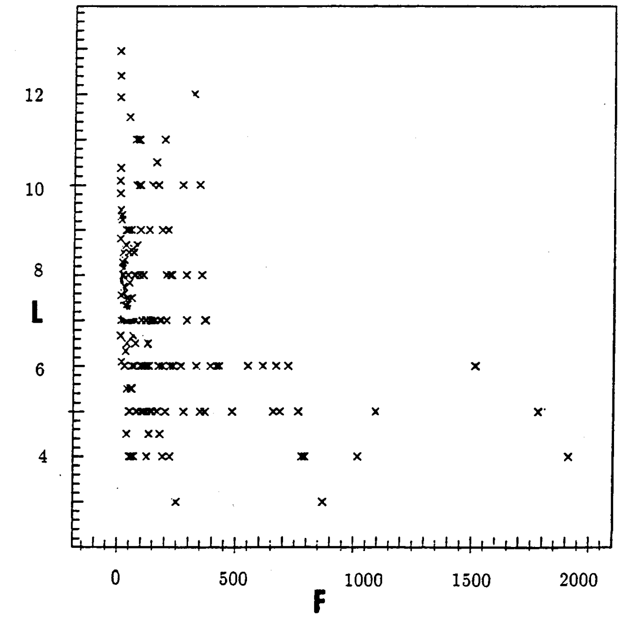


Fig. 1. Corpus data representing the dependence of word length (L) on word frequency (F); cf. Köhler (1986: 138)

Köhler did not systematically pursue this question, but re-analyzing his data, one can indeed show that with an increase of the intervals, the fit of the exponential function becomes stepwise better:

smoothing interval	R^2
none	0.24
20	0.54
50	0.77
100	0.92

Smoothing by way of moving averages, thus seems to be an effective procedure. However, Köhler was not so much interested in the fact of the gradually improving fit, as he was surprised by an oscillating curve around the theoretical hyperbolic function line: This is to say that, after using moving averages with intervals of 20, 50 and 100, a peculiar oscillation appeared which seemed to be very regular (cf. Fig. 2).

Köhler (1986) himself and, in the subsequent detail study devoted to this particular problem, Zörnig, Köhler and Brinkmöller (1990), tried to capture the course of the data by adding further power components. They thus first obtained

(1) $L = aF^b + cF^d$

and then the rather complex function

(2) $L = aF^b + cF^d + ke^{m(F-F_0)} \sin(\alpha F)$,

which captured the oscillation in a convincing way, as can be seen in Fig. 3.

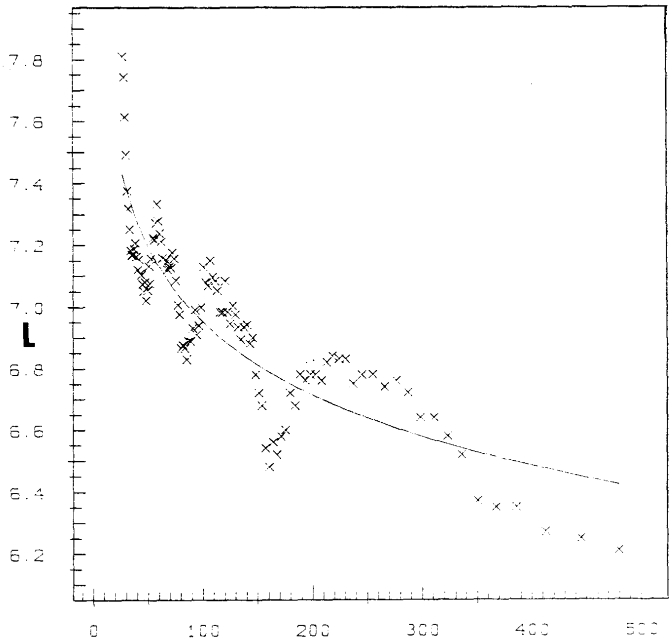


Fig. 2. Smoothing the above data (cf. Fig 1) by moving averages in intervals of 50; cf. Köhler (1986: 141)

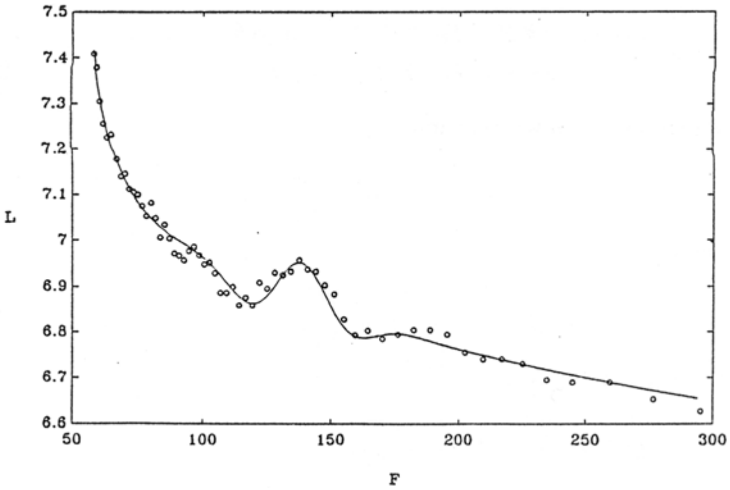


Fig. 3. Observed and computed mean lengths (Zörnig et al. 1990: 37)

What remained open, however, was a linguistic interpretation of this phenomenon, which Zörnig, Köhler, and Brinkmöller (1990: 39) left for "future research". Taking into account the complexity of formula (2), it is not really astonishing that this problem has remained unsolved until today.

In a recent study on the dependence of word length on word frequency, Strauss, Grzybek, and Altmann (2003) have examined individual texts, comparing the results to those obtained on the basis of text mixtures. In the tail of the dependence, there were many frequency classes without records, and the recorded ones contained a very small number of cases (mostly 1), thus causing a strong dispersion. Instead of smoothing the data by moving averages, they pooled low-frequency classes in order to obtain more stabile data. By pooling the data in such a way that each frequency class contained at least 10 records, the authors obtained an unequivocal corroboration of the relationship in all cases (for texts from 10 different languages)

$$L = aF^{-b} + 1 \tag{3}$$

where L = mean length, F = frequency, a and b are coefficients, and the constant 1 is the asymptote of the function (since word length was measured in terms of syllable numbers). Occurring non-syllabic words (such as, e.g., the Russian prepositions *к, с, в*), were considered as proclitics. It was not necessary to take oscillation into account, the fitting quite obviously displayed random residuals.

Irrespective of the satisfying results, it is just the observation of the lack of oscillation which again rises the question of its presence in Köhler's study; retrospectively, the problem pointed out by him remains unanswered till today, and it is not clear whether oscillation arose

(a) due to data mixing, ultimately inherent in any corpus, or
(b) as a result of an increasing sample size, or
(c) whether it was an attribute of the specific data.[2]

In the present study, an attempt shall be undertaken to offer a reason for the rise of oscillation. For our purposes, and by way of a working definition, oscillation can be assumed to be present, if the sequences of neighbouring observed data cross the theoretical curve either too frequently or too rarely. There is an interval within which the number of crossings – or the number of runs above and below the curve – can be considered to be random. We are not concerned with a time series, here, but with a sequence of numbers, capturing the length-frequency relations of words, which are ordered according to their increasing frequency. Since only low frequencies have a sufficient number of records, while higher frequencies have either none or very few records, the results for higher frequencies are insufficiently representative – therefore, great fluctuation is to be expected in this domain.

Fig. 4 illustrates the frequency-length dependence for the complete text of Puškin's verse novel *Evgenij Onegin*. Interpreting the curve, one can say that, generally speaking, high-frequency words tend to be shorter. This tendency holds true only on the average, however: whereas the curve is relatively regular at the beginning, one can observe rather irregular instabilities with the higher frequencies.

The reason for these instabilities is most likely the fact that in this particular part at the curve's tail, individual frequent words tend to have not more but one, two or three syllables. Since these words represent frequency classes in which they may occur alone, great dispersion results are to be expected exactly here. In other words: most probably, the greater dispersion is likely to be due to the insufficient number of records for these data points. Generalizing this observation, it would seem reasonable to assume that the exact part of the curve, *where* the dispersion becomes greater, depends on the distances between the represented classes, or on the frequencies of these classes (or both), which, in turn, might be related to text length (or corpus structure).

The problem at stake is obvious: if one wants to prove the validity of the law postulated by Zipf, one has three options in dealing with the data and their theoretical modelling:

[2] In another follow-up-study, Hammerl (1990) tested Polish corpus data with regard to the previously observed phenomenon of oscillation, but he did not find any.

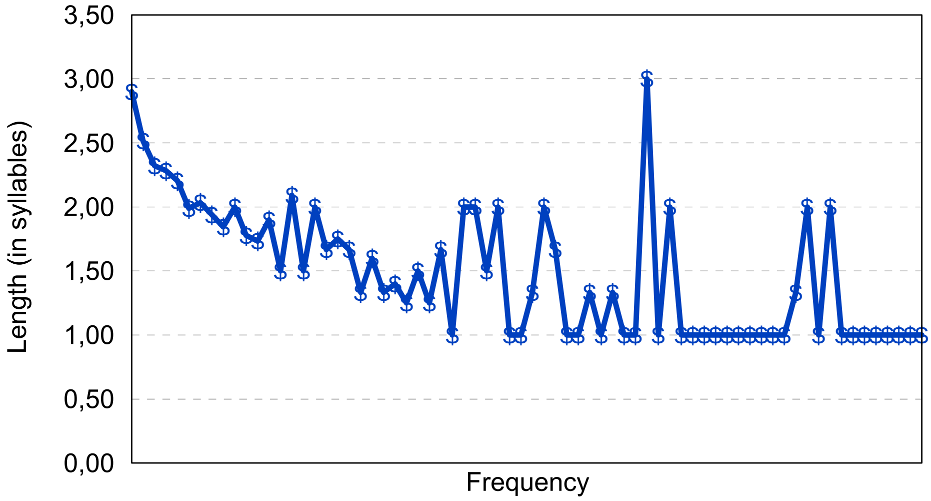


Fig. 4. The frequency-length relationship in *Evgenij Onegin*

(a) one tries to derive a curve capturing this chaotic movement, or
(b) one smoothes the data to obtain a plain course, or
(c) one smoothes the data and tries to capture also such complications (such as oscillation) which may arise during the process of smoothing.

It is obvious, that no one would ever try to go the first way, because data of this kind are extremely dispersing. Therefore, some kind of smoothing is necessary, since fitting curve (3) to the empirical data; without any smoothing, yields a very poor result (in this case $a = 2.1682$, $b = -0{,}4524$; $R^2 = 0.49$).

Now, as to the concrete manner of smoothing, different options are available: whereas Strauss, Grzybek and Altmann (2003) pooled the data and corroborated (3) in every case, another way was chosen by Köhler (1986) and Zörnig, Köhler, Brinkmöller (1990), who used moving averages and obtained the oscillating curve described above; ultimately succeeded in modelling this oscillation (see above), they had to leave open the question of its rise.

It seems most reasonable that, in one way or another, the concrete manner of smoothing is related to the phenomenon; this is not to say that oscillation necessarily is a consequence of smoothing by moving averages; yet, this might be the case in combination with a particular data structure. In an attempt to test this assumption, we will try to reproduce Köhler's finding for Puškin's *Evgenij Onegin*, and to "artificially" generate the oscillating phenomenon.

Let us start by replicating the smoothing method applied by Strauss, Grzybek, and Altmann (2003). This is to say, we first have to compute the mean length of words occurring exactly x times. For the sake of data homogeneity, we will initially concentrate on the first chapter, only; the values thus obtained are represented in Table 1 (see below). We then have to pool the data as described above, i.e. in such a way that each frequency class is based on at least ten records;[3] the resulting values of this pooling procedure are represented in Table 1.

As can be seen from Table 1, smoothing by way of pooling the classes, as described above, yields a very good result of $R^2 = 0.96$, which is graphically represented in Fig. 5.

[3] If the present results slightly differ from those presented by Strauss, Grzybek, and Altmann (2003), the reason for this is, first, that classes are pooled here "bottom-up", whereas they were pooled "top-down" in the article mentioned; and second, that the means obtained here are weighted means both for frequency and length, whereas unweighted means were calculated in the previous study.

Table 1
Fitting (3) to the data of *Evgenij Onegin*: smoothing by pooling

F	*L*	*L**
1	2.6595	2.7123
2	2.1256	1.9894
3	1.7800	1.7178
4	1.4800	1.5716
5	1.3750	1.4791
6.46	1.5385	1.3911
9.40	1.1333	1.2907
15.10	1.2000	1.1998
45.50	1.1000	1.0835
a = 1.7123, b = -0.7914, R^2 = 0.96		

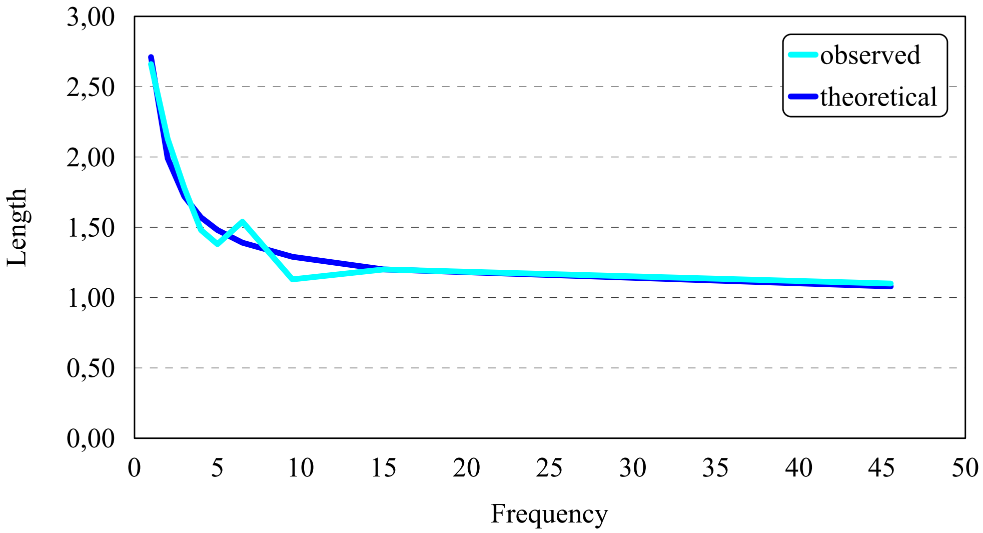


Fig. 5. Observed and computed mean lengths in *Evgenij Onegin* (ch. I)

Based on these procedures of the previous work, we may now focus the question of oscillation and extend our ruminations. Theoretically speaking, if a fitting is satisfactory, then not only small sums of squared deviations should be attained, but additionally, the empirical values should display random fluctuations around the theoretical curve. In other words: in this case, there must be neither too many nor too few runs of values on both sides of the curve. If this should still be the case, then the data either contain an intrinsic oscillation (if there are too many runs) or they display a slow wavelike motion (if there are too few runs). The oscillation must be caught by superposed curves since a simple curve cannot capture it adequately. However, if the wavelike motion arises by manipulation of data, it is not real and a simple curve is sufficient to capture it.

We can easily test this by applying the theory of runs (cf. Grotjahn 1979: 143ff., 1980). The basic idea, here, would be to test the number of sequences above and below the theoretical curve; in our case, it would be sufficient to know, if there are two few sequences (runs). In order to test this statistically, we need

n_1 number of data points above the theoretical curve (+)
n_2 number of data points below the theoretical curve (–)

r_1 number of (+)-sequences

r_2 number of (–)-sequences

$n = n_1 + n_2$

$r = r_1 + r_2$

Since we are interested in the question if there are too few runs, we test the one-sided hypothesis. As is well known, the approximation to the normal distribution may be used, for larger *n* (*n* > 30). Since the number of runs does not exceed 30, however, we have to calculate the exact (cumulated) probabilities which can also be taken from existing tables.

In order to show the rise of the wave, we first test the number of runs based on the data given in Table 2.

Table 2
Fitting (3) to the raw data in *Evgenij Onegin* (Chapter I)

F	*L*	*L**		*F*	*L*	*L**		*F*	*L*	*L**	
1	2.66	2.70	–	10	1	1.28	–	21	2	1.16	+
2	2.13	1.99	+	11	1	1.26	–	24	1	1.14	–
3	1.78	1.72	+	12	3	1.24	+	25	1	1.14	
4	1.42	1.57	–	13	1	1.23	–	32	1	1.11	
5	1.29	1.48	–	14	1	1.21	–	38	1	1.10	
6	1.43	1.42	+	15	1	1.20	–	45	1	1.09	
7	1.43	1.37	+	17	1	1.18	–	49	1	1.08	
8	1.17	1.33	–	19	1	1.17	–	68	1	1.06	
9	1.00	1.30	–	20	1	1.16	–	155	1	1.03	
$a = –0.7846$, $b = 1.6977$, $R^2 = 0.43$											

According to the description above, the positive deviations, i.e., those values which lie above the theoretical curve, are marked by (+), the negative ones, i.e., those that lie below it, by (–). Since in the whole tail, the theoretical curve lies above the empirical data, we cut off the both after the first negative sign of the last run; the number of runs thus remains constant, but the number of elements decreases thus requiring more extreme test results. As can be seen, we have

$n_1 = 6$	(6 times “+”)	$r_1 = 4$	(4 runs of “+”)
$n_2 = 14$	(14 times “–“)	$r_2 = 5$	(5 runs of “–“)
$n = n_1 + n_2 = 20$		$r_1 + r_2 = 9$	

Since the number of runs is relatively small ($r < 30$), we have to calculate the exact cumulative probability, and we thus obtain $P(R \leq r) = 0.5204$, which is not significant, of course: this is to say that the number of runs does not differ from the expected one.

Now, let us smooth the data using moving averages with larger intervals. Table 3, represents the results of smoothing with moving averages on the basis of different intervals. In the following tables interval 1 means no smoothing.

Table 3
Building moving averages and testing the runs

Interval of the moving average	n_1	n_2	n	r_1	r_2	r	$P(R < r)$
1	14	6	20	5	4	9	0.5204
2	15	5	20	4	3	7	0.2722
3	14	6	20	3	2	5	0.0173
4	12	8	20	3	2	5	0.0063

It can clearly be seen that, with an increase of the intervals, the probability for oscillation to come into play soon rises. Figure 6 convincingly illustrates this tendency, juxtaposing the results for both manners of smoothing for the sake of comparison.

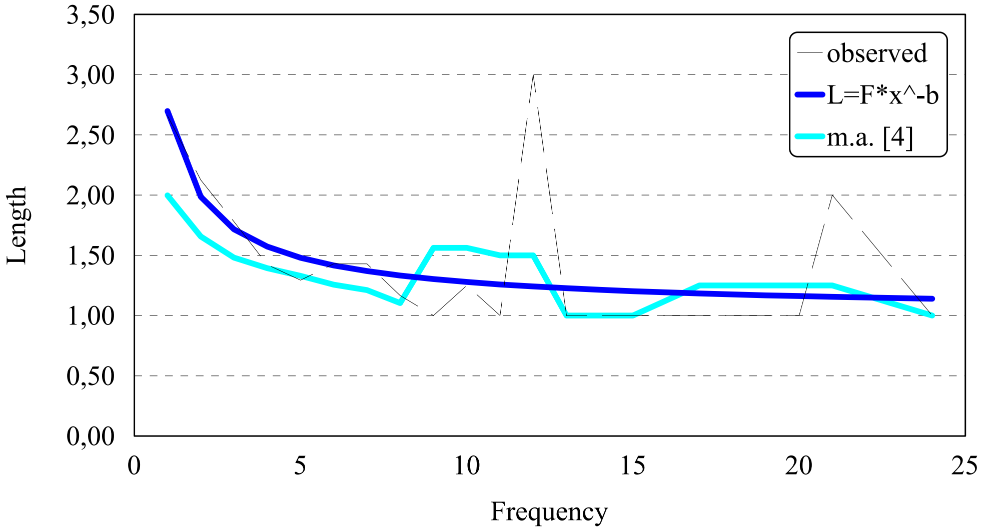


Fig. 6. Observed and computed mean lengths in *Evgenij Onegin* (ch. I)

For the sake of generalisation, let us finally extend this procedure to a broader text basis. Table 4 represents the results for each of the eight chapters of *Evgenij Onegin;* in order to eventually compare the exact results to those obtained by approximation to the normal distribution, both values are presented in parallel. The comparison of the smoothed values with the theoretical curve yields the following results (see Table 4):

Table 4
Test for the number of runs with different smoothing intervals
(*Evgenij Onegin*, chs. I–VIII)

	N	Interval	n_1	n_2	n	r_1	r_2	r	$P(R < r)$
EO 1	3086	1	15	5	20	5	4	9	0.7417
		2	14	6	20	4	3	7	0.1514
		3	14	6	20	3	2	5	0.0173
EO 2	2235	1	12	7	19	5	4	9	0.4276
		2	12	7	19	5	4	9	0.4276
		3	13	6	19	3	2	5	0.0217
		4	11	8	19	2	1	3	0.0003
EO 3	2702	1	9	7	16	3	2	5	0.0350
		2	10	6	16	3	2	5	0.0470
		3	12	4	16	2	1	3	0.0088
EO 4	2441	1	7	6	13	3	2	5	0.1212
		2	7	6	13	3	2	5	0.1212
		3	8	5	13	3	2	5	0.1515
		4	10	3	13	2	1	3	0.0455
EO 5	2310	1	10	6	16	6	5	11	0.9580
		2	10	6	16	3	2	5	0.0470
		3	9	7	16	3	2	5	0.0350
		4	11	5	16	3	2	5	0.0769
		5	11	5	16	2	1	3	0.0037
EO 6	2471	1	10	5	15	5	4	9	0.8741
		2	9	6	15	4	3	7	0.3427
		3	10	5	15	4	3	7	0.4545
		4	10	5	15	3	2	5	0.0949
		5	9	5	14	2	1	3	0.0070
EO 7	2922	1	14	8	22	6	5	11	0.5573
		2	14	8	22	4	3	7	0.0408
		3	13	9	22	4	3	7	0.0294
		4	15	7	22	3	2	5	0.0055
EO 8	3217	1	14	10	24	5	4	9	0.0857
		2	14	10	24	5	4	9	0.0857
		3	17	7	24	4	3	7	0.0450
		4	19	5	24	3	2	5	0.0209

As can clearly be seen, in most cases, intervals of three or four radically change the situation: It is almost self-evident that values of $P < 0.05$ signalize significantly few runs, i.e., the rise of a slow wave motion (with a one-sided hypothesis). As a matter of fact, by prolonging the intervals, one obtains ever longer waves, what need not be demonstrated in detail, here.

Additionally, in order to at least raise the questions of text length or data mixture, Table 5 represents the results for a successive cumulation of chapters I-VIII of *Evgenij Onegin.*

Table 5
Test for the number of runs with different smoothing intervals
(*Evgenij Onegin*, cumulated chs. I–VIII)

	N	Interval	n_1	n_2	n	r_1	r_2	r	$P(R < r)$
EO 1-2	5321	1	17	13	30	10	9	19	0.9238
		2	17	13	30	7	6	13	0.1980
		3	17	13	30	7	6	13	0.1980
		4	18	12	30	5	4	9	0.0106
EO 1-3	8023	1	21	15	36	11	10	21	0.8521
		2	21	15	36	7	6	13	0.0404
		3	25	11	36	6	5	11	0.0306
		4	27	9	36	4	3	7	0.0012
EO 1-4	10464	1	20	22	42	10	9	19	0.2204
		2	22	20	42	7	6	13	0.0036
EO 1-5	12774	1	27	21	48	14	13	27	0.8026
		2	29	19	48	10	9	19	0.0872
		3	31	17	48	7	6	13	0.0013
EO 1-6	15245	1	30	25	55	14	13	27	0.4153
		2	27	28	55	10	9	19	0.0067
EO 1-7	18167	1	35	28	63	10	9	19	0.0005
EO-tot	21401	1	33	30	63	13	12	25	0.0383
		2	32	31	63	6	5	11	0.0000

Again, one clearly sees the impact of smoothing by moving averages on the rise of oscillation; additionally, it can easily be observed that oscillation is more likely to arise for the larger, cumulated samples. If this is due to sample size, or data mixture, or to a combination of both factors, will have to be the topic of a detail study particularly devoted to this problem.

Summarizing we can state that the relation between frequency and length of words does not contain an intrinsic oscillation based on a linguistic cause, as has previously been suspected. It is simply the consequence of a special kind of smoothing.

References

Grotjahn, R. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft*. Bochum: Brockmeyer.

Grotjahn, R. (1980). The theory of runs as an instrument for research in quantitative linguistics. *Glottometrika 2*, *14-43.*

Grotjahn, R. (1992). Evaluating the adequacy of regression models: Some potential pitfalls. *Glottometrika 13*, *121-172.*

Hammerl, R. (1990). Länge – Frequenz, Länge – Rangnummer: Überprüfung von zwei lexikalischen Modellen. *Glottometrika 12, 1-24.*

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Strauss, U., Grzybek, P., Altmann, G. (2003). The more the better? Word length and word frequency. [In print]

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, Mass.: Harvard University Press.

Zipf, G.K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin.

Zörnig, P., Köhler, R., Brinkmöller, R. (1990). Differential equation models for the oscillation of the word length as a function of the frequency. *Glottometrika 12*, *25-40.*