QUANTITATIVE LINGUISTICS

Vol. 16

Studies on Zipf's Law

edited
by

H. Guiter
M. V. Arapov

Studienverlag Dr. N. Brockmeyer
Bochum 1982

QUANTITATIVE LINGUISTICS

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Studies on Zipf's law / ed. by H. Guiter; M. V. Arapov. – Bochum: Studienverlag Brockmeyer, 1982.
(Quantitative linguistics; Vol. 16)
ISBN 3-88339-244-8

NE: Guiter, Henry [Hrsg.]; GT

ISBN 3-88339-244-8

Querenburger Höhe 281, 4630 Bochum 1
Druck Thiebes GmbH & Co Kommanditgesellschaft Hagen

We would like to express our gratitude

to the

STIFTUNG VOLKSWAGENWERK

a generous grant from which made possible the translation of the Russian articles in this volume.

INHALT

$$(\text{rang} + \underline{b})^{\underline{a}} \times \text{fréquence} = \text{constante}$$

Et nous sommes convaincus que la pente a réelle n'est pas nécessairement le quotient log N/ log V" (Etienne Brunet. Le traitement des faits linguistiques et stylistiques sur ordinateur - Paris, Statistique et Linguistique, 1974, p. 111).

Nous avons eu l'occasion de montrer (Bulletin de la Société de Linguistique de Paris, 1970, 65 (I), p. 7) dans une étude portant sur six langues différentes que le produit F.R. ne demeure pas constant lorsque R augmente: "Sans nous attacher aux inévitables accidents en dents de scie, qui manifestent simplement les fluctuations aléatoires, nous constatons, sur toutes les courbes sans exception, un maximum très accusé entre les valeurs 2 et 30 de R... Un petit jeu de patience relativement facile permet d'épouser les courbes obtenues avec des superpositions de coniques. En augmentant le nombre de celles-ci, on peut s'adapter à tous les caprices".

René Michéa (Paris, Cahiers de Lexicologie, 1972, n° 20, I, p. 66) était amené à des constatations du même ordre: "Dans L'Illusion comique de Corneille, nous voyons fr croître de 946 pour le 1er mot à 3575 pour le 13ème, puis décroître pour se raccorder à une zone de moindres fluctuations".

Les statisticiens se tournèrent vers d'autres modèles rendant plus exactement compte des distributions lexicales, comme celui de Waring (XVIIIes.) appliqué à la linguistique par Gustav Herdan (1964), ou celui de Bernard Dolphin (Paris, Cahiers de Lexicologie, 1975, n° 27, I, p. 3).

Ainsi donc, i'accueil fait à la "loi de Zipf" par les linguistes français ne nous apparaît pas comme très enthousiaste.

Cette impression doit être cependant nuancée par la sagesse de certaines appréciations de René Michéa: "Mais, si la loi de Zipf procède d'une généralisation abusive, il n'en reste pas moins que la relation qu'elle implique permet de construire une base de référence indispensable pour l'étude des distributions expérimentales" (Paris, Cahiers de Lexicologie, 1973, n° 22, I, p. 73). Ou encore: "Au total, si la formule de Zipf n'est vérifiée qu'exceptionellement et approximativement, toutes les distributions expérimentales présentent en revanche, dans les limites de nos possibilités actuelles d'investigation, la même disposition hiérarchique des derniers effectifs". (Bulletin de la Société de Linguistique de Paris, 1972, 67 (I), p. 29).

La loi de Zipf n'a donc pas seulement un intérêt historique, et il apparaît justifié de revenir sur sa genèse, ses modalités et ses conséquences. Aussi, nous remercions vivement les collaborateurs qui ont bien voulu préparer les articles contenus dans le présent volume. Nous clôturons les contributions anglaises par celle (last but not least) de Bertran C. Brookes, qui nous semble porter le problème sur un plan philosophique élevé, parce qu'elle intzoduit cette question fondamentale: Pouvons-nous cerner les réalités humaines (et qu'y a-t-il de plus humain que le langage?) en leur appliquant des modèles bâtis sur l'observation du monde physique?

Henri GUITER

as all definitions are related by a power relation, $s_i = as_j^k$ with a, k constants, a > 0, where s_i and s_j are any two measures of size, and only then.

The introduction of a free parameter γ provides a degree of freedom in defining "size" in the same context, while preserving the generalized rank-size relation. In our discussion of a mathematical derivation of the rank-size relation in the context of statistical linguistics, we shall encounter a further generalization involving an additional free parameter, namely,

$$(r + m)s^{\gamma} = \text{constant}, \qquad (1)$$

where m is another constant. But now if we apply some order-preserving transformation, s → s' and expect the rank-size relation in the same form to hold, we get

$$(r + m)s^{\gamma} = \text{constant}; \quad (r + m')s'^{\gamma'} = \text{constant}$$

$$s' = ks^{\gamma/\gamma'}\left[\frac{r+m}{r+m'}\right]^{1/\gamma'} \qquad (2)$$

Unless m = m', s' is not a function of s alone, as we would expect from any reasonable definition of size. It depends also on rank, that is, implies a bizarre or, at any rate, an involved redefinition.

As stated, Zipf thought that he derived his rank-size relation from a general law governing human behaviour, which he called The Principle of Least Effort (Zipf, 1949). Since the derivation led to a mathematical relation between variables, one might suppose that it was carried out by mathematical deduction from mathematically stated postulates based on operational definitions of in principle measurable quantities. This was not the case. Zipf introduced the Principle of Least Effort in a discussion involving words that could be interpreted in various ways without any one interpretation fitting all the situations in which the principle was supposed to apply.

Underlying the statement of the principle is an assumption to the effect that people faced with a choice among alternative courses of action will choose the course that minimizes "work" in some more or less long-term sense. In each case, however, "work" retains at most its connotative meaning, which is insufficient as a starting point of a mathematical derivation.

To cite an example given by Zipf, suppose two towns separated by a mountain range are to be connected by a road. Possible alternatives are (1) to build a road directly up and down the range; (2) to build a road with smaller gradients over the nearest mountain pass; (3) to bore a tunnel through the range. Method (3) might minimize work involved in traveling but would cost much work in construction. Method (2) would save work in construction but would make a longer road. Method (1) would shorten the distance but would mean lifting loads to higher levels. Zipf's definition of "effort" mentions all these considerations. He assumes that whatever solution is adopted minimizes "effort" over some "long run". That is, he invokes some statistical expectation of "work" over some stretch of the future. How this horizon and the probabilities entering the expectations are to be estimated remains obscure, and a usable definition of "work" is left to the reader's imagination.

In short, the principle of Least Effort is stated in vague, connotation-ridden language, precluding rigorous deduction. This is not to say that intuitive notions or arguments stated in every day language have no place in scientific discussion; only that a derivation of a mathematically stated relation on the basis of verbal arguments suggests a lack of understanding of what a mathematical approach to a science involves and what it takes to justify it.

Be this as it may, let us follow Zipf's "derivation" of the rank-size relation, rs = constant, in the context of statistical linguistics, the relation usually associated with Zipf's Law.

In every language, many words are carriers of several meanings. The more meanings are assigned to single words, the more the vocabulary of a language can be compressed. From the point of

view of the speaker, therefore, "effort" would be minimized if words were loaded with very many meanings, for then he would need to acquire only a small vocabulary to express his thoughts. This tendency to compress the vocabulary by the use of words heavily loaded with meanings Zipf calls the Force of Unification. From the point of view of the hearer, on the other hand, "effort" is minimized if each word carries exactly one meaning because the hearer exerts effort in deciphering the meaning of the words used by the speaker. It is, therefore, in the interest of the hearer to expand the vocabulary. Zipf calls this tendency the Force of Diversification. The actual distribution of meanings among words appears to be a compromise between the interest of the speaker and that of the hearer. Further, the distribution of meanings among words must be related somehow to the relative frequencies with which the words appear in the language. This is the point of departure for Zipf's argument.

Let the most frequently occurring word occur with frequency F_1. If this word carries m_1 meanings, and if the average frequency of occurrence of each of these m_1 meanings in discourse is f_1, we have

$$m_1 f_1 = F_1 \qquad (3)$$

So far, so good. Equation (3) is, in fact, a tautology, inasmuch as the average frequency of occurrence of meanings associated with the most frequent word must be F_1/m_1. Here, however, Zipf becomes obscure. He writes:

> "Obviously, the Force of Unification which theoretically acts in the direction of putting all different meanings behind a single word will tend to increase the size of m_1 at the expense of f_1. On the other hand, the Force of Diversification which theoretically acts in the direction of reducing the number of different meanings per word will tend to increase f_1 at the expense of m_1. Therefore the respective sizes of m_1 and f_1... will represent the action of opposing forces of Unification and Diversification". (Zipf, 1949, p. 28).

Since relation (3), in fact $m_r f_r = F_r$, is a tautology, "opposing forces" have nothing to do with establishing it. What Zipf needs for his subsequent argument is the relation $f_r = km_r$ for all words of rank r (r = 1, 2, ...) ranked according to their frequency of occurrence. He simply states this relation, in fact with k = 1, assuming it to be a consequence of the "opposing forces":

> "However if $m_1 = f_1$ and since $m_1 \times f_1 = F_1$, then clearly m_1 will be the square root of F_1..." (p. 28).

The rank-size relation, rs = constant, can be derived from $f_r = km_r$, if a further assumption is made that the number of meanings assigned to a word is inversely proportional to the square root of the rank. For $m_r = k'r^{-1/2}$ and $f_r = km_r$ together imply $F_r = Kr^{-1}$, where $K = kk'^2$. It remains to identify the "size" of a word, s, with its frequency of occurrence.

The relation $m_r \sim r^{-1/2}$ is fairly well corroborated empirically in English corpuses. The English Semantic Count lists 20.000 words ranked by frequency of occurrence. When the average number of meanings per word in groups of 1000 are plotted against the median ranks of the groups, i.e., the 500th, the 1500th, ... etc. on log-log paper, a graph is obtained, fitted very well by a straight line with slope -0.4656 ± 0.0027, which is close to -0.5, implied by the relation $m_r \sim r^{-1/2}$.

Thus, because the relation $m_r f_r = F_r$ is a tautology, the rank-size relation associated with Zipf's Law is a consequence of the empirically observed (approximate) relation $m_r = kr^{-1/2}$ and of an ad hoc assumption that the frequency of occurrence of a meaning associated with a word is proportional to the number of meanings associated with that word. This is, of course, a very strong assumption, and its connection with a "balance of opposing forces" is difficult to see.

In view of the above analysis, we feel justified in ignoring Zipf's attempt to derive the rank-size relation in statistical linguistics from theoretical considerations. We will examine similar attempts in other contexts below. For the present, let us turn to evidence for Zipf's Law in language statistics.

Studies of rank-size relations in language corpuses antedate Zipf's investigations. J.B. Estoup noted the reciprocal rank-size

relation in 1916. E.V. Condon (1928) noted that word counts by L.P. Ayres (1915) and by G. Dewey (1923) corresponded to (rank) x(frequency) = constant. The first example of this relation cited in Zipf's Human Behavior and the Principle of Least Effort was based on word counts by Hanley (1937) in James Joyce's novel Ulysses, which contains 260,430 running words, of which 29,899 are different and ranked. The rank-size graph on log-log paper is compared with that for a sample of American newspapers (43,989 running words, 6002 different). The linear relation, $\log r + \log s = \text{constant}$ is well corroborated by both graphs. The positions of the two parallel lines are governed by the sizes of the respective corpuses. Many other similar graphs are presented in Zipf's book.

Interest in rank-frequency relations has been sustained. Subsequent studies revealed marked and consistent deviations from Zipf's Law at the high frequency end of the scale. Moreover, the slope of the linear part of the plots was frequently observed to be different from -1, that is, better fitted by the formula $rs^{\gamma} = \text{constant}$. We shall return to this point.

Although Zipf's explanation of the rank-frequency relation on the basis of references to the Principle of Least Effort are anything but convincing, they contain a germ of an idea, which was subsequently developed with full mathematical rigour by B. Mandelbrot (1953). Mandelbrot derived the rank-size relation for language corpuses from some theoretical considerations. The relation turned out to be a further generalization of Zipf's Law, and it accounts for some systematic deviations from the relation $rs^{\gamma} = \text{constant}$. The basic idea in Mandelbrot's formulation is minimization under constraint, for which he used an appropriate mathematical apparatus.

The idea in Zipf's discussion developed by Mandelbrot is that the speaker, in attempting to minimize his "effort" must do so under a constraint, namely, the necessity to imbue his utterances with information, which in Mandelbrot's model appears in its mathematically defined sense.

Mandelbrot identifies the effort in producing a word with its length. Length, of course, can be defined in several ways, e.g.,

as the time required to produce the word, the number of letters in its written form, the number of phonemes identified in its spoken form, etc. To fix ideas, Mandelbrot takes the number of letters. He speaks of "cost" rather than of effort, and we will follow his usage.

Imagine that we undertake to invent a language, more precisely a vocabulary of R words. At our disposal is an alphabet of G letters. We are free to attach several meanings to a word, and we will assume that the words with the larger semantic load will occur with larger frequencies in our language. From the point of view of a speaker attempting to minimize the average cost per word, it would be most advantageous to assign all possible meanings to a single word -- the cheapest. Then all his utterances would consist in repetitions of that word. But his utterances would convey no information, because the hearer would know in advance what the speaker was going to say. In a moment, we will show that this conclusion follows from the mathematical definition of the amount of information.

Thus, while the speaker would indeed be minimizing the cost of his utterances if he confined all of them to the cheapest word, the purpose of communication would be defeated. Communication is served only if words carry some minimum average amount of information.

Suppose now we are given a repertoire (or a source, as it is called in information theory) of R signals, of which the r-th signal appears in a message with probability P_r. Assuming statistical independence among the occurring signals, the average amount of information per signal in any message issuing from the source is defined by

$$H = -\sum_{r=1}^{R} p_r \operatorname{Log} p_r. \qquad (4)$$

The base of the logarithm merely determines the unit of information; so without loss of generality we can take the logarithm to the natural base e. It is easily seen that H is maximized if $P_r = 1/R$ (r = 1, 2, ... R). Also H tends to zero as any of the

P_r tends to 1 (hence all the others tend to zero). This corroborates our contention that if every message is confined to a single signal, the communication becomes devoid of information. Since this would defeat the purpose of the speaker (to communicate), his attempt to minimize cost must be constrained by the requirement that a certain minimum positive average amount of information per signal is maintained.

In view of this constraint, the speaker will have to use more than one word in his utterances. He is free, however, to distribute the meanings he wishes to convey in any way he likes among the words. According to our assumption above, this will introduce a distribution of frequencies with which the words will occur, namely the word of rank r will occur with some relative frequency (or probability) P_r.

Let the cost (proportional to the number of letters) of the r-th ranking word be C_r. Then the average cost of a word in the language so constructed by the speaker will be

$$C = \sum_{r=1}^{R} C_r P_r \qquad (5)$$

The Principle of Least Effort can now be seen in the formulation of the following problem:

$$\text{Minimize} \quad \sum_{r=1}^{R} C_r P_r \qquad (6)$$

$$\text{Subject to} \quad -\Sigma P_r \log P_r = H > 0 \qquad (7)$$

$$\Sigma P_r = 1; \quad P_r \geq 0 \quad (r = 1, 2, \ldots R) \qquad (8)$$

The objective function (6) to be minimized is the average cost per word expressed as a function of the relative frequencies of occurrence of the words in the language. Constraint (7) expresses the requirement that the average amount of information

per word must be positive. Constraint (8) expresses the fact that the P_r, being probabilities, must be non-negative and must sum to unity.

The minimization problem can be solved with the aid of Lagrange multipliers. To do this, we form the expression

$$-\Sigma P_r \log P_r - \beta \Sigma P_r C_r + \alpha \Sigma P_r, \qquad (9)$$

where α and β, the Lagrange multipliers are constants to be determined later, and set the partial derivatives with respect to each P_r (r = 1, 2, ... R) equal to zero. We obtain

$$-1 - \log P_r - \beta C_r + \alpha = 0 \quad (r = 1, 2, \ldots R) \qquad (10)$$

Solving for each P_r, we obtain

$$P_r = \text{Exp}\{(\alpha - 1)\} \text{Exp}\{-\beta C_r\}. \qquad (11)$$

Next, we undertake to determine C_r as a function of r. If the average cost per word (i.e., the average number of letters) is to be minimized, we should utilize all the letters of our alphabet to make single-letter (shortest) words. Of these there will be G. Then. we will utilize all ordered pairs of letters, of which we shall have G^2, then all ordered triples, etc. Moreover, we shall want the shortest words to be the most frequent. Thus, words of rank 1, 2, ... G will be one-letter words, words of rank G+1, G+2, ... $G+G^2$ will be two-letter words; and, in general, words of ranks from $G^{r-1}+G^{r-2}+\ldots+1$ to $G^r+G^{r-1}+\ldots G$ will be r letters long. Fixing the unit of cost as the cost of one letter, we find that the r-th word in our arrangement from the cheapest (and therefore the most frequent) to the most expensive (and least frequent), assuming arbitrary assignment of rank to words of equal length, will cost approximately $\log_G r$. That is, proportional to $\log_e r$, where the constant of proportionality is determined by the cost unit. Substituting this relation into (11), we obtain

$$rP_r^{\gamma} = \text{constant}, \tag{12}$$

where the constant and the parameter γ depend on H and on the Lagrange multipliers α and β. These are determined so that the constraints (7) and (8) are satisfied.

We see that (12) has the form of (generalized) Zipf's Law, where P_r, being the relative frequency of occurrence of a word in a corpus, represents "size". Writing (12) as

$$P_r = Kr^{-1/\gamma} \tag{13}$$

we see that if the vocabulary of our language is "infinite", we must have $\gamma < 1$ to insure the convergence of ΣP_r to unity. Also this point will be discussed below.

A further generalization can be obtained by dropping the assumption that all the letters of our alphabet cost the same. Let the costs of the letters be $C_1, C_2, \ldots C_G$. Then the number of words of cost C is determined by the following difference equation:

$$N(C) = N(C-C_1) + N(C-C_2) + \ldots N(C-C_G). \tag{14}$$

On the right side of (14), the i-th term represents the number of words of cost C in which the first letter is C_i. Therefore N(C), the sum of these terms represents the number of words of cost C.

Let M be the largest root of the equation

$$\sum_{i=1}^{G} M^{-C_i} = 1. \tag{15}$$

Then, if we again arrange the words in terms of decreasing cost, C_r will be given to a first approximation by

$$C_r = \log_M r. \tag{16}$$

Thus, again C_r is proportional to $\log_e r$, and we obtain the rank-size relation (13). But if we use the next approximation for C_r in solving the difference equation (14), we obtain

$$C_r = \log_M(r + m) + C_o \tag{17}$$

where m and C_o are constants. Choosing appropriate units and substituting (17) into (11), we obtain

$$(r + m)P_r^{\gamma} = \text{constant}, \tag{18}$$

which is the generalization of Zipf's Law mentioned above. Mandelbrot writes Equation (18) as

$$P_r = A(r + m)^{-B}, \tag{19}$$

with A, m, B constants. Clearly B corresponds to $1/\gamma$ in our Equation (18). In an infinite corpus, r ranges from 1 to infinity. In this case, P_r is a proper probability distribution only if $B > 1$. The restriction does not apply to finite corpuses.

Mandelbrot's mathematically respectable formulation of the germinal idea contained in the Principle of Least Effort raises the question of whether regularities observed in rank-size relations in other contexts can be similarly derived. Examination of Zipf's attempts to do so by concatenations of vaguely formulated assumptions, occasionally illustrated by a simple equation, are of no help. One example will suffice.

The classical rank-size relation often holds for cities of a country ranked by population. One such relation is shown in Table 1.

In attempting to account for this relation, Zipf identifies work involved in the production of commodities (L) and in transporting them (M) with effort. Since M is positively related to the distance (D) over which the commodities are transported, it follows (so the argument goes) that as M decreases it becomes more economical to transport goods to the consumer, whereas if

Table 1. Rank-size relation for largest U.S. cities (census of 1960)

City	Rank	Population	(Rank)×(Population)×10^{-6}
New York	1	7.710.300	7.7
Chicago	2	3.511.000	7.0
Los Angeles	3	2.450.000	7.4
Philadelphia	4	1.971.200	7.8
Detroit	5	1.654.100	8.3
Houston	6	932.600	5.6
Baltimore	7	922.200	6.5
Cleveland	8	869.100	7.0
St. Louis	9	747.100	6.7
Milwaukee	10	732.600	7.3
S. Francisco	11	716.300	7.9
Dallas	12	672.400	8.1

M increases, it becomes more economical for the consumer to travel to the centers of production. It is not clear why this should be so, since transportation costs can be expected to be associated also with moving people as well as commodities. Nevertheless, Zipf again sees the Force of Unification (concentration of people) opposing the Force of Diversification (dispersion of people). As the forces are "balanced", the effort $(m+L)_h$, associated with the production and transportation of commodity h will determine the distance D over which this commodity will be traded for other commodities. As the radius of this trading area decreases, the area will decrease as the square of the radius. Hence the number of different domains where h is produced will be proportional to $(M+L)^2$. But as the amount of h, F_h, allotted to a domain will be inversely proportional to the number of domains, Zipf feels justified in writing

$$F_h(M + L)_h^2 = \text{constant}. \qquad (20)$$

From here on, Zipf has smooth sailing. All he needs to do is to introduce mediating variables to substitute for F_h and for $(M+L)_h$ and to invoke the "balance" between the Force of Unification and the Force of Diversification to arrive at

$$rP_r^\gamma = \text{constant}, \qquad (21)$$

where r is the population rank of a city or a metropolitan area and P_r its population.

The ad hoc nature of Zipf's assumptions piled on top of each other reveals the weakness of his approach. On the other hand, his drive to seek some unifying principle underlying an impressive number of observed regularities in an impressively wide range of contexts (of which we have mentioned only two) is readily understandable in the light of contemporary sociology of knowledge. Physical science, unified by the deductive power of mathematics, has fulfilled Francis Bacon's prophecy, "Knowledge is power". The prospect of extending man's power over nature to power over "himself" (i.e., over social dynamics and institutions) nurtures hope in the power-oriented and apprehension in those who perceive this extension of power as a threat. The former see in "unified social science" an instrument of control over human masses. Some (though by no means all) of the latter see "unified social science" as a source of understanding, which might enable people to resist the hegemony of power. Wishful thinking becomes incorporated in epistemological predilections of both groups. It is clearly discernible in Zipf's tortuous reasoning. Zipf was apparently deeply impressed by the theoretical power of the Principle of Least Action in physics and was deluded into believing that he found an analogue in his Principle of Least Effort, which bears a verbal resemblance to the former.

Had Zipf paid more attention to the abundance of rank-size relations in contexts totally unrelated to "economy of human effort", he might have had second thoughts on the subject. On the other hand, even this is doubtful, since he subsumed the distribution of the numbers of species among genera under the same principle. The genesis of a statistical regularity in a stochastic process seems not to have occurred to him. Had it occurred, he might have found himself in the position of Laplace, who, when asked by Napoleon why there was no mention of the Creator in his treatise on celestial mechanics, replied, "Sire, I had no need

of that hypothesis". On the other hand, it may have been as difficult for Zipf to abandon his belief in the Principle as it is for many to abandon the belief in a Creator.

In what follows, we will examine the statistical approach to rank-size relations. Here observed regularities will be derived with mathematical rigour (weakened only by the use of approximations) as asymptotic distributions generated by stochastic processes. This approach, like that of classical mathematical physics, is also endowed with power of theoretical unification, but this power, unlike that of classical physics, is not rooted in the universality of virtually unassailable physical laws. Rather, it stems from structural isomorphisms between the axiomatic bases of stochastic processes, isomorphisms that are totally unrelated to the content of these processes.

Consider the immense range of contents in which the normal distribution is observed. It fits well the distribution of sizes of beans and to the distribution of I.Q. scores in a large human population. But it would be foolhardy to conclude that the connecting link is the circumstance that people eat beans. The pervasiveness of the normal distribution is a consequence of the central limit theorem. Roughly, if we have a population of things that suffer increments or decrements in size independent of their sizes, this stochastic process will generate an equilibrium distribution, which is normal. Again, if the increments or decrements suffered by our things are proportional to the sizes already attained, the asymptotic distribution will be logarithmic normal. And so on for many different kinds of distributions.

The distributions just mentioned are size-number rather than rank-size distributions, but each can be derived from the other. In particular, if (rank)×(frequency) = constant, where f is the frequency of occurrence of a word in a corpus, then $n(f)f^2$ = constant, where f is the frequency of occurrence of a word in a corpus, and n(f) is the number of different words that occur with frequency f. We will use this relation to derive Zipf's Law from a stochastic process.

In spite of the fact that the statistical approach makes no reference to a "law of nature", let alone to any specific principle

of "minimized effort", there is a connection between the derivation of an asymptotic distribution and a minimization (or maximization) procedure. As we have seen, Mandelbrot's derivation of the rank-frequency relation involves a minimization problem. On the face of it, the derivation is not based on an underlying stochastic process. Nevertheless, a tacitly assumed stochastic process lurks in the derivation. Note that Mandelbrot's minimization problem has a dual maximization problem: maximize the average information per word under the constraint that the average cost per word is kept to some specified minimum. Now information, as it is defined in this problem, is structurally isomorphic to entropy. The problem, then, amounts to that of maximizing the entropy of some system under a constraint. Problems of this sort arise in statistical mechanics. For example, given a gas confined to a volume and isolated from the environment (a closed system), it is desired to derive the distribution of velocities of the molecules that constitute the gas. Changes in the velocities of single molecules, resulting from impacts, are regarded as a stochastic process. The ultimate (equilibrium) distribution is the asymptotic distribution of this stochastic process. It is derived via maximizing the entropy of the gas.

Now the maximization of entropy in a closed system is regarded as a "law of nature". Nevertheless, this law (the Second Law of Thermodynamics) has been shown to be rooted in statistical "laws". In fact, according to one philosophical view, all the "laws of nature" are in the last analysis rooted in statistical "laws", and the strict deterministic nature of these laws is only a consequence of the law of large numbers, which holds with great exactness for objects with which classical physics has been dealing. The validity of this view is not our concern here. It does, however, seem reasonable to assume that when science attempts to deal with phenomena that cannot be related to the unassailable physical laws, the way to proceed is not via inventing additional "laws" of supposed comparable validity and generality but rather to fall back on statistical analysis, whereby regularities involving at least very large populations can be expected to be derived.

We will now examine H. Simon's (1955) derivation of frequency-number relation in statistical linguistics, of which, as has been said, the rank-frequency relation is an immediate consequence. It is well to guard against possible confusion between two meanings of "frequency". In the context of statistical distributions, "frequency" usually refers to a magnitude proportional to the probability of observing a member of a population characterized by a value of a random variable in a given range. Thus, the magnitude of a random variable is the "independent variable" and frequency is the dependent variable. In our discussion, frequency is the "independent variable" of a statistical distribution. It refers to the frequency of occurrence of a given word in a corpus. The dependent variable is the number of different words in the corpus that occur in the courpus with the corresponding frequency; or, if this number is divided by the number of different words in the corpus, the fraction of the different words that occur with the corresponding frequency, in other words an estimate of the probability with which a word of a given frequency ("size") is selected.

Consider the process of producing a corpus, say an author writing a book. The act of writing can be conceived as that of selecting successive words out of one's repertoire of words (vocabulary). This model may not correspond to the author's mental processes. As I write these lines, I do not feel that I grope for each successive word. The thoughts come to me in larger units. But in constructing a model of a stochastic process, one must start somewhere "to fix ideas". Let us therefore assume that choices of successive words are the elementary probabilistically determined events and see where this assumption will lead us.

At every moment of decision, a choice is influenced by two factors, which Simon calls "association" and "imitation". Association is the influence that stems from the author's concern of what he is writing about, thus with the tendency to choose words that have already occurred in the corpus. For instance, as I write this, the words "corpus", "stochastic", "frequency", etc. have higher probabilities of being chosen than, say, "castle" or

"winery", which would have larger probabilities of being chosen if I were writing a tourist's guide of Austria. Imitation is the influence of usage in the language in which the author is writing. For instance, certain words are bound to occur in almost any English corpus, regardless of subject matter, because they are incorporated in the structure of English, e.g., "have", "will", "because", etc. Saussure's distinction between "la parole" and "la langue" reflects these factors. The former refers to the character of the verbal output of a particular speaker or writer; the latter to the character of the language as a whole.

Consider now the accumulation of words in a corpus as it is being produced. Suppose k words have already been written. Let n(f,k) represent the number of <u>different</u> words that have already appeared f times. Then fn(f,k) is the total number of occurrences of these words (counting repetitions). Simon assumes that under the influence of association, the probability that the next word selected will be one that has already occurred f times is proportional to fn(f,k). That is to say, with respect to the frequency of occurrence of the words in category f (words that have appeared f times), the corpus of k words is a representative sample of <u>these words</u> in the author's vocabulary, or, more precisely, of the portion of his vocabulary that is relevant to the subject matter of the corpus.

Turning to imitation, we can assume that a word selected under that influence, which has already occurred f times in the corpus, has a probability of less than f/k of being selected as the next, i.e. (k+1)st, word, because unlike association, imitation may introduce new words (which have not yet appeared) into the corpus, simply because they are in the language. Moreover, we may suppose that this reduction of the probability of choosing a word that has already appeared f times will be greater when f is small than when f is large, because words in the latter category can be assumed to be the "heavy duty" words, commonly occurring in any corpus and thus occurring with frequencies approaching those in the corpus independent of content. Accordingly, Simon assumes

that the probability that a word in category f, chosen on the basis of imitation will be chosen with probability $(f-c)n(f)/k$ $(0<c<1)$.

If words are chosen by imitation with probability q and by association with probability (1-q), then the joint probability that a word is chosen by imitation and belongs to category f will be given by $q(f-c)n(f,k)/k$. Similarly, the joint probability that a word will be chosen by association and will belong to category f will be given by $(1-q)fn(f,k)/k$. Adding the joint probabilities, summing over f, and subtracting from 1, we have the probability that the (k+1)st word will not belong to any category f, that is will be a new word: $\alpha(k) = qc(n_k/k)$, where n_k is the total number of different words in k running words.

Simon's model of the accumulating corpus can now be formulated as the following difference equation:

$$n(f, k+1) - n(f,k) = K(k).(f-1)n(f-1,k) - K(k)fn(f,k) \quad (f = 2,3,\ldots,k)$$

$$n(1, k+1) - n(1,k) = \alpha(k) - K(k)n(1,k) \qquad (22)$$

The first terms on the left are, strictly speaking, expectations rather than actual numbers of words in category f. Since, however, we are interested in the equilibrium distribution resulting from this stochastic process, we can treat the expectations as if they were actual numbers. The factor of proportionality, K(k), although independent of f, must be supposed to depend on k. The first term on the right of the first equation represents a gain for category f when the (k+1)st word has been added, if that word belongs to category (f-1). The second term represents a loss for category f, if the next word belongs to that category, since its addition results in the "promotion" of this word to category (f+1). Thus, the difference represents the expected gain (positive or negative) for category f. Similar reasoning leads to the second equation referring to category 1. On the assumption that in an accumulating corpus the author's vocabulary becomes gradually depleted, we can suppose that the probability

of adding new words, $\alpha(k)$ is a monotone decreasing function of k. One of Simon's models is developed on the basis of this assumption. In what follows, we present his simplest model, where α is assumed to be independent of k.

The factor K(k) can be determined from the following considerations. Since $K(k)fn(f,k)$ is the probability that the (k+1)st word belongs to category f, that is, <u>has</u> occurred in the corpus of k words, summing over f must yield the probability that it is <u>not</u> a new word. Therefore

$$\Sigma K(k)fn(f,k) = K(k)\Sigma fn(f,k) = 1 - \alpha. \qquad (23)$$

On the other hand, we have

$$\Sigma\ fn(f,k) = k, \qquad (24)$$

the total number of running words in the corpus. Combining (23) and (24), we have

$$K(k) = (1 - \alpha)/k. \qquad (25)$$

When equilibrium is established, all the frequencies in the corpus grow proportionately. Thus, for any two categories f and (f-1), we must have

$$\frac{n(f,k)}{n(f-1,k)} = \frac{n(f,k+1)}{n(f-1,k+1)} = \beta(f) \qquad (26)$$

where $\beta(f)$ is independent of k. Also the <u>relative</u> numbers of words in the several categories, $n^*(f,k)$ will be independent of k, and we can write $n^*(f,k) = n^*(f)$. We proceed to calculate $\beta(f)$.

In view of equations (22), (25), and (26), we can write

$$\left[\frac{k+1}{k} - 1\right]n(f,k) = \frac{1-\alpha}{k}\left[\frac{f-1}{\beta(f)} - f\right]n(f,k). \qquad (27)$$

Solving for $\beta(f)$ in terms of f, we obtain

$$\beta(f) = \frac{(1-\alpha)(f-1)}{1+(1-\alpha)f} = \frac{n^*(f)}{n^*(f-1)} \quad (f = 2,\ldots,k) \tag{28}$$

By iterating the relation $n^*(f) = \beta(f)n^*(f-1)$ and denoting $(1-\alpha)^{-1}$ by ρ, we obtain

$$n^*(f) = \frac{(f-1)(f-2)\ldots 2.1}{(f+\rho)(f+\rho-1)\ldots(2+\rho)}\, n^*(1) \quad (f = 2,\ldots,k) \tag{29}$$

To determine $n^*(1)$, we note that the total number of different words in a corpus of size k is

$$n_k = \sum_f n(f,k) = \alpha k = \alpha \Sigma f n(f,k). \tag{30}$$

By reasoning used in deriving (29), we obtain

$$\left[\frac{k+1}{k} - 1\right] n(k) = \alpha - \left[\frac{1-\alpha}{k}\right] n(1)$$

$$n(1) = k\alpha/(2-\alpha) = n_k/(2-\alpha)$$

$$n^*(1) = n(1)/n_k = (2-\alpha)^{-1} = \rho/(\rho+1) \tag{31}$$

Now (29) can be written as

$$\frac{\Gamma(f)\,\Gamma(\rho+2)}{\Gamma(f+\rho+1)} \cdot \frac{\rho}{\rho+1} = \frac{\Gamma(f)(\rho+1)\,\Gamma(\rho+1)\rho}{\Gamma(f+\rho+1)(\rho+1)}$$

$$= \frac{\Gamma(f)\,\Gamma(\rho+1)\rho}{\Gamma(f+\rho+1)}$$

$$= \rho B(f,\rho+1) = n^*(f) \tag{32}$$

Thus, (31) and (32) give the desired probability distribution for words of frequency f.

Let us now compare the probability distribution derived by Simon with the rank-size relation expressed by Zipf's Law. Suppose α is very small, so that ρ is very close to 1. Then $B(f,\rho+1) \cong B(f,2) = \Gamma(f)/\Gamma(f+2) = 1/f(f+1)$. Consider now the sum $\sum_{f\cong s}^{\infty} [f(f+1)]^{-1}$, which for large values of s can be approximated by

$$\int_s^\infty \frac{df}{f(f+1)} = \lim_{t\to\infty} \int_s^t \left[\frac{1}{f} - \frac{1}{f+1}\right] df = \log\left[\frac{f}{f+1}\right]_s^\infty =$$

$$= \log\left[\frac{s+1}{s}\right] = \log\left[1 + \frac{1}{s}\right] \cong \frac{1}{s} \tag{33}$$

But $\sum_{f\cong s}^{\infty} [f(f+1)]^{-1}$ is proportional to the number of different words that have appeared at least s times in the (very large) corpus. This number corresponds approximately to the rank of the words that has appeared s times. Hence rank is approximately inversely proportional to "size" s, which is the original formulation of Zipf's Law.

The approximation of the sum by the integral is not justified for small values of f. In fact, for $\rho = 1$, we have

$$B(1,\rho+1) \cong B(1,2) = 1/2;\ B(2,2) = 1/6;\ B(3,2) = 1/12 \tag{34}$$

That is to say, about half of different words in a large corpus are supposed to appear just once, about 1/6 just twice, about 1/12 just three times, etc. These are the low frequency words with large values of r. Since, however, successive ranks are assigned to words with the same frequency, the reciprocal rank-frequency relation does not apply to these words. At the high frequency end, Simon's model predicts the (approximately) reciprocal rank-frequency relation for $\rho \cong 1$, but it does not account for the discrepancies typically observed at this end, namely a "flattening out" of the rank-frequency line on log-log paper.

Turning to Mandelbrot's formula $P_r \sim (r+m)^{-B}$, we see that it does account at least qualitatively for the discrepancy through the additional parameter $m > 0$, which, we recall, Mandelbrot derived from his cost minimization model. On the other hand, Mandelbrot's model has little to say about the low frequency end (where r is large), again because words with the same (low) frequencies are assigned different ranks.

In sum, although the analytic forms of the two models are similar, they differ essentially both in their underlying assumptions and in their predictions at the two extremes of the frequency scale. In principle, therefore, the two models could be pitted against each other in empirical contexts. This was one of the issues in several exchanges between Simon and Mandelbrot over some years following the publication of Simon's paper on skew distributions (Mandelbrot, 1959, Simon, 1960, Mandelbrot, 1961a, 1961b, Simon, 1961a, 1961b, 1963).

Mandelbrot points out (1959) that with very few exceptions, his parameter B (cf. Eq. (19)) turns out to be greater than 1, whereas Simon's parameter ρ, which corresponds essentially to the reciprocal of B must be greater than 1 if α is constant. Simon (1960) disputes Mandelbrot's contention, pointing out that in data presented by Zipf, ρ turns out to be greater than 1 more often than less than one and cites many examples. Thus, on the basis of the estimated magnitudes of the exponential parameter, the issue between the two models cannot be decided. Questions of convergence of the terms of the probability distributions are hardly relevant in empirical contexts, because no corpus is infinite. As stated above, the restriction on B to values greater than 1 does not apply to finite corpuses; nor does the restriction $\rho>1$. To be sure, if α is constant and positive, then ρ must be greater than 1 by definition. However, the analogous parameter can be smaller than 1 in variants of Simon's model where $\alpha(k)$ decreases monotonically with k or else vanishes for k larger than some k_o.

In his next paper, Mandelbrot (1961a) sidesteps the issue of the magnitude of the exponential parameter. Instead, he states

the central issue to be the question of whether the entire class of frequency-number distributions considered by Simon can be subsumed under a single generating scheme, as Simon attempts to do in his 1955 paper. The general form of these distributions is (in our notation)

$$n(f,k) = K(k)f^{-(\rho+1)}, \qquad (35)$$

where n is the number of items that occur with frequency f in a corpus of size k, and K(k) is some function of k. It would, of course, be fortunate if all such distributions could be derived from some very general principle of generation in the way that the normal distribution is derived from random increments of "size" suffered by a large population of items, where the increments are independent of the size already attained. In Mandelbrot's opinion, Simon did not succeed in this task. In particular, Mandelbrot argues that Simon's "heuristic" arguments used in deriving similar distributions when the rate of appearance of new items, $\alpha(k)$ is not constant, do not stand up to rigorous analysis. To my knowledge, the last paper pertaining to this controversy was published by Simon (1963). In it, results of a large number of Monte Carlo simulations based on Simon's models are presented. They are offered as evidence that the stochastic processes considered by Simon do generate the distributions associated with Zipf's Law. Actually two separate issues seem to have been involved in the exchange. One was related to the empirical adequacy of Simon's models; the other to the mathematical rigour of his derivations. The first issue is difficult to decide, because available data are not usually precise enough to distinguish between small differences between results derived from the different models and because the "best" estimation procedures of parameters are often unknown (something that Simon pointed out in one of his papers). The purely mathematical issues are another matter. The arguments and counter-arguments of both authors on this score are involved, and further discussion would not fall within the scope of this paper. The interested reader is referred to the abovementioned papers by Mandelbrot and Simon.

Whether or not the "Zipf distributions" (as Mandelbrot calls them) can be derived from a particular stochastic model, such as that proposed by Simon, it is intriguing to speculate whether some more general type of stochastic process underlies all or most of them, for instance a "birth and death" process, where items enter the collection and disappear from the collection in accordance with specified probabilities that depend on the distribution of "sizes". Simon attempted to interpret his model in this way. In the context of a language corpus, he considers a "slice" of the corpus, where the leading end moves as in the cumulating corpus, adding words, while words are dropped from the trailing end. In this scheme, Simon assumes that if one representative of a particular word is dropped, then all the representatives of that word are dropped, suggesting that this assumption is approximately justified, "if the representatives of each word instead of being distributed randomly through the sequence, were closely 'bunched'" (Simon, 1955, p. 432). The assumption is clearly a very strong one.

Birth-and-death stochastic processes were classified by Horvath and Foster (1963) in accordance with the probabilities governing gains and losses. Let the elements of our population be groups of individuals and the size of a group be identified by the number of individuals in it. Groups grow when individuals enter them and shrink when individuals leave them. In particular, a group may disappear from the collection when all the individuals comprising it leave. A 2 x 2 classification of these processes is shown in Table 2. The entries are the equilibrium probability distribution of group sizes.

Simon derived Fisher's logarithmic series in one of the variants of his model. The exponential and the truncated Poisson distributions are not of the type considered here. Of special interest is the Yule distribution, a member of the "Zipf-Distributions" with $\rho<1$. It was derived by G. Yule (1924) from a stochastic model proposed to explain the distribution of biological genera comprising a given number of species. Yule assumed that the probability of a mutation producing a new species occurring in a particular genus during a given short interval of time is proportional to the number of species in that genus, while the probability of a generic mutation (producing a new genus) is proportional to the number of genera. Thus the entry of new species (individuals) occurs with probability proportional to the size of the group (number of species in a genus), while the "death" of a species occurs when its genus disappears (and with it all the species in that genus).

Table 2. Distribution associated with various probabilities governing the growth and depletion of groups.

		Probability of joining a group: Independent of group size	Probability of joining a group: Proportional to group size
Depletion: Only through complete dissolution of groups		Exponential distribution	Yule distribution
Depletion: Through departure of individual members		Truncated Poisson distribution	Fisher's logarithmic series

More recently, W.H. Horvath and C.C. Foster (1963) proposed the same sort of model to account for the distribution of sizes of war alliances. (An alliance may be thought of as an analogue to a biological genus and the nations in it as analogues to species). In this model, nations enter the "international system" partitioned into alliances and either join an already existing alliance or constitute "singlet" alliances. Once in an alliance, a nation cannot leave it except when the whole alliance breaks up (as it usually does following a defeat or frequently also following a victory). Not surprisingly, the model leads to a distribution of sizes of alliances analogous to Yule's distribution of sizes of genera.

In his studies on "deadly quarrels", Lewis F. Richardson (1960) examined the distribution of sizes of Chicago gangs in the 1920's and sizes of Manchurian marauder bands in the 1930's (the

latter inferred from casualties inflicted in raids). He found that both distributions were fitted by the same formula with nearly equal parameters, which led him to conjecture that "organization for aggression" was an underlying principle common to both distributions. This conjecture may seem as far-fetched as Zipf's Principle of Least Effort. Nevertheless, like Zipf's Principle in the context of word frequency distributions, it may contain a small kernel of truth. Namely, the stochastic process underlying this type of distribution is based on the assumption that while new members are recruited into groups (with probabilities proportional to the sizes of the groups), a member cannot leave a group except when the group breaks up. This is in accord with the observation that it is very difficult to "resign" from a gang. In contrast, stochastic processes where units can freely leave clusters as well as join them lead to size distributions of quite different types. (The interested reader is referred to studies on sizes of casually forming groups reported by J. Coleman (1964).)

In sum, "explanations" of the so called Zipf's Law by postulated underlying stochastic processes point to a unifying principle, which may play a role in the social sciences analogous so that played by "natural laws" in the physical sciences. The entire theoretical edifice of physical science rests on a solid foundation of such virtually unassailable laws, e.g., the matter-energy conservation laws, laws governing the propagation of electro-magnetic waves, the law of universal gravitation, etc. No such comparable laws are known that could serve as an analogous foundation for social science. However, many social phenomena may well be describable in terms of stochastic processes. Given sufficiently massive bodies of homogeneous data, so that stability of statistical regularities can be expected, and fine distinctions between different types of processes can be reflected, a typology of stochastic processes may provide a methodologically fruitful classification of social processes. The approach carries promise for constructing "unified" theories of social phenomena, unified not in the sense of similarity of content but rather in the sense

of a structural similarity of underlying stochasticalla formulated dynamics. We have seen that this interpretation of Zipf's Law has yielded considerable theoretical dividends.

REFERENCES

AYRES, L.P., A Measuring Scale for Ability in Spelling. New York, 1915

COLEMAN, J.S., Introduction to Mathematical Sociology. London, 1964

CONDON, E.V., Statistics of vocabulary. Science, 67, 1923, 300

DEWEY, G., Relative Frequency of English Speech Sounds. Cambridge, Mass., 1923

ESTOUP, J.B., Gammes Stenographiques. Paris, 1916

HANLEY, M.L., Word Index to James Joyce's "Ulysses". Madison, Wis., 1937

HORVATH, W.J. & FOSTER, C.C., Stochastic models of war alliances. Journal of Conflict Resolution, 7 (1963), 110-116

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

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

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

MANDELBROT, B., (b) Post scriptum to "Final Note". Information and Control, 4 (1961), 224-228

RICHARDSON, L.F., Statistics of Deadly Quarrels. Pittsburgh, 1960

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

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

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

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

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

YULE, G.U., A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Philosophical Transactions, B, 213 (1924), 21

ZIPF, G.K., Human Behavior and the Principle of Least Effort. Reading, Mass., 1949

A Variational Approach to Frequency-Rank Distributions of Text Elements

M.V. Arapov, Moscow

0. Introduction

0.1. The overall objective of this paper is to develop a better understanding of the role which probability plays in natural language. We consider the two basic notions of this paper, namely 'text' and an 'articulation of the text' as undefinable.

Neither of the notions cited seems necessary for formulating our final results, neverless we need to explain both why we want to obtain these results and why we consider them interesting.

In present-day linguistic usage the word 'text' has two different meanings: A philologist regards a text as a language form by means of which its author expresses his relatively coherent vision of the world using means that are specific to a given culture and age. Under this interpretation, a text is an event in a chain of events that make up the history of a particular culture, while philology itself can be regarded as a chain of reflexions on the sense of this culture.

A linguist dedicated to quantitative studies of language interprets the same concept of the text differently: every passage of words seems to be the text for him. He implements the standard tools of statistical analysis to this passage regardless of its cultural value.

From the philological point of view, an integral text of a novel story, poem, scientific or journal paper etc. and sometimes their independent parts are genuine texts, whereas a telephone directory, dictionary or random sample from a genuine text are examples of quasi-texts. Quasi-texts account for a great part of language material (Saussure's language), whereas the genuine texts are - metaphorically speaking - only small diamond crystals in an amorphous mass.

The text as an integral entity can be articulated into elements of various types: paragraphs, sentences, words, morphemes, phonemes etc. It may require deep penetration into the author's concept of a text to arrive at the proper articulation of the text and, as a rule, the articulation obtained is not the only one possible. Let us suppose some definite articulation of the text has been chosen already; let us assume it is an articulation into words.

The most puzzling fact about such an articulation is that the distinct elements of the same text always have sharply differentiated frequencies of occurrences. Existing theory does not offer any satisfactory account of this differentiation, still it provides some means to describe the variation of the observed frequencies. A conventional device is frequency-rank distribution where the frequency F_n is treated as the function of rank n, that is, the number n of the text elements whose frequencies $F \geq F_n$. There are empirical formulas e.g. Zipf's Law for the word frequencies or the geometrical distribution for the phoneme frequencies by which the functions F_n can be satisfactorily approximated.

To put it differently, the form of frequency-rank distributions remains unchanged (up to the value of its parameters) when one text is replaced by another. Nevertheless, in the course of such replacement, the frequency of an individual element may change considerably.

0.2. We believe the basic facts about word frequency distribution are well known to readers of this article. We want only to turn their attention to one detail: these facts have been generally treated in the framework of statistical linguistics, one of the possible methods of interpreting and using the probability theory and mathematical statistics for the description of a language. If one utilizes statistical linguistics, one adopts a paradigm that regards some questions as feasible and others as meaningless. It is quite reasonable from the viewpoint of statistical linguistics, for instance, to average characteristics of the text element such as its frequency over large corpora of texts, totally discarding any individualities of the texts examined.

In the present article we try to outline an alternative paradigm (see also Arapov et al. 1975, 1975a). Our general idea is that the individual text, if it is a genuine one (the text from philologist's point of view), has a specific structure which manifests itself in quantitative ratios that hold for the results of the articulation of the text and subsequent classification of its elements. To systematize and interpret these quantitative ratios, we propose to use the variation principle: of all conceivable classifications one which is realized is that for which the probability of realization is maximal. The probability of realization must be defined over some appropriate space of all classifications admissible under given constraints. Each classification is at the same time a distribution of elements singled out among the cells, so it is possible to speak of the most probable vector of quantities - numbers of the elements which refer to each cell.

The proposed approach is not absolutely new in the field of quantitative linguistics (see Mandelbrot 1951, Šrejder 1967, Levič 1978) but its novelty would fail to be noticed, if the principles on which this approach was based were not confronted with underlying ideas of statistical linguistics.

1. The Statistical Linguistics: Basic Concepts

Statistical linguistics rests upon the following hypothesis.

> Each language unit x has an invariant probability p_x of occurence in the corpus of texts X. (1)

How tenable is this hypothesis? We cannot observe the probability of any event directly but only through its frequency. The relation between the frequency F_x of x and the probability is given by the law of large numbers.

Let L be length of the text X (the term length applies here to the total number of word occurrences of the text). If $L \to \infty$,

then the probability of the following inequality tends to unity,

$$|Lp_x - F_x| < \varepsilon \qquad (2)$$

where ε can be chosen arbitrarily small.

Since the inequality (2) by itself gives no indication as to the method of testing the hypothesis (1), we need some additional postulate about the process of text construction. The simplest assumption that is usually advanced is as follows:

> Text generation is a sampling process, i.e. a sequence of independent realizations of the same random variable U. (3)

The values that U assumes are words $x_1, x_2, \ldots$ which together form vocabulary V. Let us denote by $p_1, p_2, \ldots$ the probabilities of the corresponding values $x_1, x_2 \ldots$

Having adopted model (3), we can set up many criteria for the statistical homogeneity of the text. One such criterion is as follows. Let us divide the text being examined into two parts, each with length L_1 and L_2, $L_1 + L_2 = L$, and calculate F_1 and F_2 - the numbers of occurrences of x in both parts of the text. If the probability of usage of the element x in text X was stable (i.e. the postulate (1) holds), then the difference $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1 = \frac{F_1}{L_1}$ and $\hat{p}_2 = \frac{F_2}{L_2}$, would be small. Postulate (3) enables us to evaluate the variance $s^2_{\hat{p}_1 - \hat{p}_2}$ of $\hat{p}_1 - \hat{p}_2$ and thereby to calculate the probability of observed value of $\hat{p}_1 - \hat{p}_2$.

Since the distribution of the statistics

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}_1 - \hat{p}_2}} \qquad (4)$$

where $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{L_1} + \frac{1}{L_2}\right)}$, $\hat{p}L = F_1 + F_2$, can be approximated by the standardized normal one, then the probability that $z > z_o$ would be $1 - \Phi(z_o)$,

$$\Phi(z) = \frac{2}{\sqrt{2\Pi}} \int_0^z e^{-\frac{t^2}{2}}\, dt.$$

As a numerical example we now examine the usage of one of the most frequent Russian words - the preposition v 'in'. There are as many as 15 935 occurrences of this word in the plays that have been included in the corpus of Zasorina's Russian Frequency Dictionary (1977), the total length L_1 of the dramatic texts examined is 287 000 running words. The total length L_2 of all texts from newspapers and magazines is 251 000 words and the number of occurrences of v is 9 630.

$$\frac{15\,935 - 9\,630}{\sqrt{\left(\frac{15\,935 + 9\,630}{287\,000 + 251\,000}\right)\left(1 - \frac{15\,935 + 9\,630}{287\,000 + 251\,000}\right)\left(\frac{1}{287\,000} + \frac{1}{251\,000}\right)}} = 29.4.$$

Hence the observed value of the difference $\hat{p}_1 - \hat{p}_2$ is about 30 times greater than its standard deviation. The probability of this value is negliglibly small ($<10^{-100}$). Therefore the hypothesis about the existence of the invariant probability p_x for the two parts of the corpus should be rejected. The same holds for other parts of the corpus, and the result obtained is typical of many words.

However, test (4) and other similar tests do not prove that hypothesis (1) is false. (What (4) really discredits is the conjunction of (1) and (3), the postulates which together constitute the theoretical base of the existing method of evaluating the reliability of a frequency dictionary, see Frumkina 1964).

At least two strategies of defense for hypothesis (1) have already been suggested; both, of course, require us to admit some additional assumptions about the text structure.

Using the first strategy, we can suppose that there are not one but many probabilities $p_x^{(i)}$ of using x in separate parts $X_1, X_2, \ldots, X_i \ldots$ of the corpus $X = \bigcup_i X_i$; these parts are occasionally termed sublanguages.

A shortcoming of this strategy is that it enquires a very loose concept of 'sublanguage'. The limits of an individual sublanguage cannot be established a priori and the search for homogeneity may lead to successive splitting of the corpus into smaller and smaller parts without achieving stability of the probabilities, because the narrow subject of the assembled texts by itself does not guarantee the stability of probabilities in the texts, as has been demonstrated in the information retrieval studies.

The natural limit of the splitting seems to be an individual finite text, but the very concept of probability becomes meaningless for such a text, because of its finiteness (we cannot implement (2) here).

Using the second strategy, we can introduce a more complex model than that of text generation (3). Let us assume, for instance, that the probabilities are assigned not to the individual text elements but to pairs of them, i.e. elementary events are now elements of V^2. As a first step, we pick out a pair $\langle x_i x_j \rangle$ with some probability p_{ij}, then, in the second step, one of pairs $\langle x_j x_k \rangle$ with probability $\frac{p_{jk}}{\sum_n p_{jn}}$ and so on (Markov process).

If the probabilities of the pairs satisfy some simple conditions, then assumption (1) holds true automatically. On the other hand, in the case of the Markov process, the variance $s^2_{\hat{p}_1-\hat{p}_2}$ cannot be evaluated by such a simple method as above. Since the realizations of the random variable U are not independent any more, the total variance is not the sum of L variations of the separate trials. It is not difficult to build up an analogue of test (4) for the text regarded as a realization of a Markov process: still, it is impossible to carry out necessary computations without detailed description of the process itself. But as a rule, an investigator is not in a position to gather all the necessary quantitative data to obtain such detailed description of the text generation process.

All attempts to reject the approach based on (1) have led to a stalemate so far: whenever a "trap" such as test (4) is constructed, it can be avoided by making the model of text generation

more sophisticated. The more sophisticated the model is, the more complicated experimentum crucis must be. Such an experiment necessarily involves a lot of intricate calculations whose accuracy depends on the soundness of many intermediate hypotheses proposed in order to compensate for the lack of empirical data or to avoid mathematical difficulties.

The paradigm of statistical linguistics cannot be proven to be inadequate, but still we need to depart from this paradigm if we want to ask some naive questions. One of such questions has been formulated above: Why do the word frequencies within any text scatter so widely?

From the point of view of statistical linguistics, this question is equivalent to the question of why the probabilities $p_1, p_2, \ldots$ of elements $x_1, x_2, \ldots$ are different. As remarked above, the occurrences of the elements $x_1, x_2, \ldots$ are treated in this case as elementary events, whereas contemporary probability theory rules out the question of origin for probabilities of elementary events.

However, this theory does not demand that the space of elementary events should be introduced exactly in such manner as has been done in statistical linguistics. In the next section, we shall examine a totally different way to do that.

2. The Classifications and Related Statistics

2.0. Accepting hypothesis (1) in one form or another, statistical linguistics as a theory attempts to solve two not necessarily related problems at the same time. One is to establish ratios of the word frequencies within a single text, the second, to find out an invariant of the word usage in an assembly of texts (in a general population) - the probability p_x of element x.

The importance of the first problem in a hierarchy of problems treated by statistical linguistics is minor, though one acknowledges its importance formally. The theory simply states that for some obscure reason the probabilities of word occur-

rences p_x can be presented as a decreasing series, the general term of which p_n, has the form:

$$p_n = \frac{C}{n^\gamma} \qquad \text{(Zipf's Law)} \qquad (5)$$

or

$$p_n = \frac{C_1}{(n + B)^\gamma} \qquad \text{(Zipf-Mandelbrot's Law).} \qquad (5')$$

If we accept the hypothesis (3), the hyperbolic series (5) of probabilities p_n induces a similar distribution of the observed frequencies F_n: the similarity must rise with the size of the sample. The hypothesis of the hyperbolic distribution is not entirely supported by empirical evidence: for large samples the distribution of the word frequencies does not generally conform to Zipf's Law. Apart from this, the acceptance of (5) or (5') as the probability of an elementary event leads to some mathematical difficulties: an infinite probability space can be determined by a series, if and only if the series converges. The series (5) or (5') converges only if $\gamma>1$, but in real samples the value of γ can be less than unity (see Arapov 1978).

The intrinsic logic of statistical linguistics does not imply (5) or any function of that sort (besides (5), some authors suggest the negative binomial, lognormal, Weibull's and other distributions); this hypothesis is completely independent of other postulates of the theory.

Both the first and the second problem posed by statistical linguistics can be solved outside the paradigm. To solve the second, some probabilistic invariant of word usage must be introduced, but not necessarily in form (1), cf. Arapov et al. 1978.

2.1. General outline of the explanation of the frequency scattering proposed by us was given earlier. We suggested that the observed distribution of the word occurrences among cells of a certain classification had to be the most probable one under certain constraints.

Let us consider this approach more closely. Let two sets be given, the set X = {x} of N elements and set Y = {y} of m cells. To find the most probable distribution of N elements among m cells, we must define which distributions are admissible and which of them are different. There are many ways to do so, but we are interested in ways that yield non-trivial distributions.

In all the examples of classifications listed below, it is assumed that

i. Each element x of the set X belongs to a cell y_i;
ii. There is no element that belongs to two or more cells.
iii. There are no empty cells.
iv. The elements x and cells y are enumerated by natural numbers from 1 to N or m respectively (but not so in Example 4 below).

Conditions i-iii induce some equivalence relations in X. Condition iv introduces the relation of linear order into the sets X and Y separately. (A relation of another sort will be defined instead of linear order in Example 4).

By a tradition which has its origin in statistical physics, the different methods to enumerate all conceivable variants of element distribution among some cells are called statistics.

Example 1. (Maxwell-Bolzmann's statistics)[1]

Let two partitions Y_1 and Y_2 of the set X into m subsets (cells) be equivalent if one can be transformed into another by changing the numbers of the elements $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$ belonging to the same cell y_i.

If Y contains the cell y_i with n elements, then there are at least as many as n! partitions Y' which are equivalent to Y. Let an ordered m-tuple $n_1, n_2, \ldots, n_m$, $\Sigma n_i = N$, give the numbers of elements that belong to the cells $y_1, y_2, \ldots, y_m$ of the partition Y (such m-tuple of numbers we term cell frequencies, henceforth CF). Then the overall number Φ of the partitions which are equivalent to Y can be obtained by the following formula:

$$(N; n_1, n_2, \ldots, n_m) = \prod_{i=1}^{m} n_i!$$

The class of equivalent partitions is termed configuration. To transform one configuration into another, it is necessary to transfer at least one element from one cell to another.

The following simple combinatorial formula gives the number of different configurations with fixed CF which cannot be transformed one into another by permuting the elements within a cell:

$$S_1(N;\ n_1,n_2,\ldots,n_m) = \frac{N!}{n_1!n_2!\ldots n_m!} \qquad (6)$$

In statistical physics the logarithm of S_1 multiplied by some irrelevant constant is called entropy.

If the equivalence of the partitions is established, then (under fixed N and m) only CF determines the shape of the distribution. Let us define the probability of a given CF. As an elementary event we take a single configuration and set the probability of each configuration equal to $1/\Sigma S_1$, the summation in the denominator is over all m-tuples $n_1,n_2,\ldots,n_m$, $\Sigma n_i = N$ (i.e. over all CF).

The probability of a distribution is the sum of the probabilities of all configurations with the same CF, that is, the product of the number of configurations $S_1(N;\ n_1,n_2,\ldots,n_m)$ and the probability of a single configuration $1/\Sigma S_1$.

Finally, the problem of finding the most probable distribution reduces itself to that of obtaining CF which renders the functional (6) a maximum.

Example 2. Let two partitions Y_1 and Y_2 of the set X into m cells be equivalent if one can be transformed into another either by

a) changing the numbers $x_{i_1},x_{i_2},\ldots,x_{i_n}$ of the elements which belong to the same cell y_i, or

b) changing the numbers $y_{j_1},y_{j_2},\ldots,y_{j_s}$ of s cells which contain equal quantities of the elements.

In this case, the class of the equivalent partitions can be defined by means of a sequence of numbers a_k, where a_k is the quantity of the cells with exactly k, k = 1,2,3,... elements;

a_k must satisfy the following conditions:

$$\sum_{k=1}^{N} k a_k = N \qquad (7)$$

$$\sum_{k=1}^{N} a_k = m \qquad (8)$$

If the sequence a_k is given, we can immediately get corresponding CF: if $a_k = r$, then there are exactly r numbers n, n=k in CF.

The number of distinct configurations can be obtained by the formula:

$$S_2(N;\ a_1,a_2,\ldots,a_N) = \frac{N!}{(1!)^{a_1}a_1!\ldots(N!)^{a_N}a_n!} \cdot \qquad (9)$$

It is only natural to set the probability of each distinct configuration equal to $1/\Sigma S_2$; the summation in the denominator is over all sequences a_k which satisfy the conditions (7) and (8). Then the probability of a given sequence a_k (and corresponding CF) is the product of the number of distinct configurations $S_2(N;\ a_1,a_2,\ldots,a_N)$ with a given sequence a_n and the probability of a single configuration $1/\Sigma S_2$. By direct computation one can find that under the statistics of the Examples 1 and 2, the probabilities of the same CF are different.

Ju.K. Krylov has suggested the following modification of the statistics examined: the cells with equal numbers of elements are still indistinguishable, although the order of the elements within a given cell is relevant (condition "a" does not hold).

Example 3. (Bose-Einstein's statistics).

In both examples cited, having exchanged an element of one cell for an element of another, we get a new configuration with the same CF. (In Example 2 it occurs only if such transfer is not

equivalent to the permutation of some cells). Now consider statistics under which two partitions of the set X are equivalent if one can be obtained from another by arbitrary permutations of the elements of X. Then the partitions can differ from one another only by CF, and each CF corresponds to only one configuration.

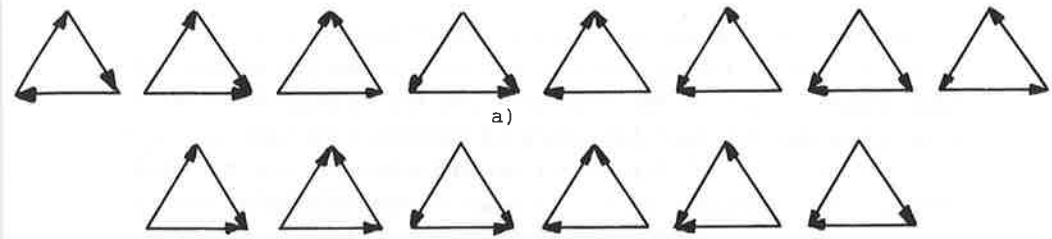
Hence the overall number of ways to distribute N elements among m cells is now equivalent to that of presenting the quantity N as the sum of m positive summands, their order being relevant and none of them being equal to zero. This number is $\binom{N-1}{m-1}$ and the probability of a distinct configuration is naturally 1/ $\binom{N-1}{m-1}$.

Example 4. There is only one detail which distinguishes this statistics from Maxwell-Bolzmann's (see Example 1). For this statistics (and for statistics of the examples 2 and 3), we suppose that there is established a linear order of the elements in the set X, that is, the antisymmetrical, antireflexive and transitive binary relation which holds for every pair $\{x,x'\}$ from X^2.

Let us now consider the relation which has all these properties save one, namely, transitivity. There is no special term for such a relation, but a graph by means of which it can be depicted is called a tournament, therefore we term the relation t-order (see Harary, Palmer 1973). All eight t-orders (tournaments) which can be established in a set of three elements are shown in Fig. 1a. They can be compared with six graphs corresponding to all possible ways to define the 'normal' linear order in the same set. It is easy to verify that there may be as many as $2^{\binom{n}{2}}$ different t-orders in the set which consists of n elements.

Then we introduce the new statistics in exactly the same way as we did in Example 1. We consider two partitions to be equivalent if they differ one from another only by the t-order of the elements within the same cell. The order of the cells themselves is regarded as linear.

The overall number of distinct configurations yields the following formula,



a)

b)

Fig. 1.

$$S_4(N;\ n_1,n_2,\ldots,n_m) = \frac{2^{\binom{N}{2}}}{2^{\binom{n_1}{2}}2^{\binom{n_2}{2}}\ldots 2^{\binom{n_m}{2}}} \qquad (10)$$

The space of elementary events and the probability of a given CF can be introduced exactly as in Example 1.

2.2. Let some statistics be chosen. Then we are in a position to solve the variational problem which was formulated earlier, that is, to find the CF rendering the chosen functional a maximum. If the statistics of Examples 1,2 or 4 are chosen it can be the functionals (6), (9), or (10) respectively.

In the case of Bose's statistics (Example 3), the solutions are necessarily trivial, because every CF has the same probability. The examples cited do not exhaust all possibilities, of course.

The values of the functionals (6), (9), and (10) differ from the corresponding probabilities by a proportionality factor which is equal to the probability of a single configuration, but this factor is inessential for the solution of the variational problem. The problem itself can be solved in one of two modifications: (i) the number of the cells is treated as given in advance (we suggested such an approach earlier but primarily for the sake of simplicity), or (ii) the number of cells is limited only by condition $m \leq N$. Since the two approaches differ by the proportionality factor, transition from the first to the second does not alter basically the results obtained.

However, the obtained solution itself turns out to be a disappointingly trivial one (for all statistics listed and for both modifications of the problem): for the Maxwell-Bolzmann's statistics and for the statistics of Example 4 we have (i) $n_i = N/m$, or (ii) $n_i = 1$ for all i; as to Example 2, the simplest result can be obtained when N is a so called triangular number, that is, when it can be represented as a sum $N = 1 + 2 + 3 + \ldots + r$. In case that all $a_k = 1$, therefore the cell frequencies (CF) form the series of natural numbers: 1,2,3,...,r. (It holds for Krylov's modification of this statistics also). If N is not a triangular number, then there will be some repetitions of natural numbers in the series or some omissions of them.

In statistical physics, it is additional constraints that are responsible for the non-triviality of the obtained distributions. The constraints originate in some principles of conservation. Consider the way in which such constraints work.

Let a set of N elements be distributed among m cells by the procedure known from Example 1 (Maxwell-Bolzmann's statistics), iε be a quantity assigned to each element of the cell y_i, $1 \leq i \leq m$. We want to find the most likely distribution of the cell frequencies (CF) under the following additional conditions

$$\sum_{i=1}^{m} i\varepsilon n_i = \varepsilon \sum in_i = E \qquad (11)$$

$$\sum n_i = N$$

where E has fixed value (A.P. Lević suggested the term "limiting factor" for such a requirement, see Lević, 1978). For the sake of simplicity, let $\frac{\varepsilon N}{m} \Sigma i > E$, so that a uniform distribution of CF is excluded in advance.

A certain distribution of the elements is shown in Fig. 2. Suppose this distribution is the most probable, that is, it

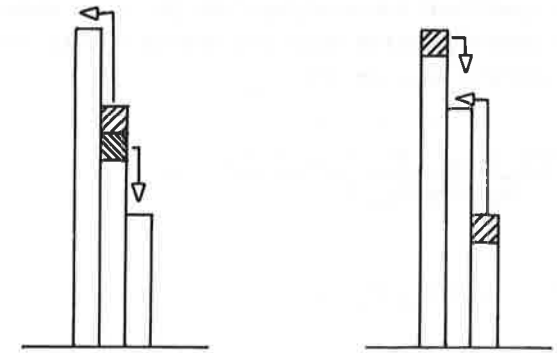


Fig. 2.

renders (6) maximum and the conditions (11) are fulfilled. Let us find the ratio of n_{i-1}, n_i, and n_{i+1}, i.e. the function n_i which realizes a maximum of the functional (6).

Requirement (11) makes it impossible to transfer one element from one cell to another, but does not exclude the possibility of taking two elements from the cell y_i and placing one of them into the cell y_{i-1} and the second, into the cell y_{i+1}. Such transfer does not change the value of E, because an additional element being placed into cell y_{i-1}, increases E by ε, and another being placed into cell y_{i+1}, decreases E by the same ε.

According to our assumption, the value of the functional (6) is the greatest possible for a given configuration, and the value of the product $n_{i-1}!n_i!n_{i+1}!$ must be the smallest possible, so that the transformation shown in Fig. 2a can only make the product greater. Thus,

$$\frac{(n_i - 2)!(n_{i-1} + 1)!(n_{i+1} + 1)!}{n_{i-1}!n_i!n_{i+1}!} \geq 1,$$

whence

$$n_{i-1}\, n_{i+1} \geq n_i^2 - n_{i-1} - n_i - n_{i+1} - 1 . \qquad (12)$$

Next let us transform the distribution as it is shown in Fig. 2b (such a transformation does not change E either); then reasoning in a similar way, we get

$$\frac{(n_i + 2)!(n_{i-1} - 1)!(n_{i+1} - 1)!}{n_{i-1}!n_i!n_{i+1}!} \geq 1,$$

whence

$$n_i^2 + 3n_i + 2 \geq n_{i-1}n_{i+1}. \qquad (13)$$

Combining (12) with (13), we arrive at the following double inequality:

$$n_{i-1}n_{i+1} + n_{i-1} + n_i + n_{i+1} + 1 \geq n_i^2 \geq n_{i+1}n_{i-1} - 3n_i - 2 .$$

Dropping the linear terms of the inequality, we get the following proportionality,

$$\frac{n_{i-1}}{n_i} = \frac{n_i}{n_{i+1}}$$

That is, the cell frequencies form geometric series and in general are not equal to one another, although the partition follows Maxwell-Bolzmann's statistics as before.

The derivation of Bolzmann's distribution of number n_i of particles with a given energy is discussed above is in fact only a simplification of the standard derivation of this law in physics (though similar distributions occur in linguistics and biological taxonomy also, see Levič 1978).

The underlying idea of the derivation cited is that of the limiting factor (11), and the same principle was used in some works on Zipf's Law (Mandelbrot 1951, Śrejder 1967, Levič 1978). (To obtain the hyperbolic distribution this principle must be applied in a more subtle way, but we shall not go into that here.)

Sometimes the limiting factors can be introduced naturally enough, but often one has to invent such factors ad hoc. They have

to be specific for a given subject field, though the distributions in these fields are similar and the procedures of classification are identical. One may or may not regard the special nature of the limiting factors as a shortcoming of the approach depending on one's epistemological views, yet it is tempting to avoid use of any such factors in the derivation.

2.3. As it turns out, this is possible if not one but two coordinated partitions Y and Y* of the set X are examined simultaneously.

A partition Y* of the set X is called co-partition with respect to the partition Y of the same set X, if two conditions are fulfilled:

(i) for any cell $y_i \in Y$ and any $y_j^* \in Y^*$ either their intersection is empty or it contains exactly one element;

(ii) if $y_i \cap y_i^*$ is not empty, then for every y_k such that $n_k \geq n_i$ the intersection $y_k \cap y_j^*$ is not empty either.

Let us take another look at the definition of co-partition. Let $y_1, y_2, \ldots, y_m$ be cells of Y. Let us enumerate the elements which belong to every cell by natural numbers $k = 1,2,3,\ldots,p$.

Now we can set up the cells of the co-partition Y*, placing the elements with the numbers i into the cell y_j^*. The equivalence of both definitions of co-partition is obvious.

REMARK. It is clear that p is equal to the number of elements in the largest cell; if the cells are arranged in order of decreasing volume, then $p = n_1^2$.

If the cells of both the partitions Y and Y* are arranged in order of decreasing size, we can represent the partitions by means of a diagram (see Fig. 3). Elements x of the diagram that have no neighbours above and to the right of them are called limit elements (they are marked out in Fig. 3). For every limit element x, the numbers n and m of the elements in cells y and y*, $x \in y \cap y^*$ can be treated as a sort of Cartesian coordinates of x. The problem of finding such cell frequencies of Y and Y* which are most likely is equivalent to that of finding the most probable coordinates n and m of the limit elements.

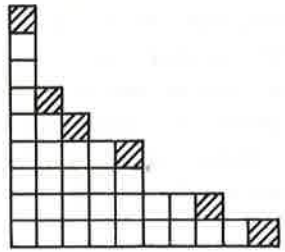


Fig. 3.

It follows from what we have discussed above that the solution is determined by the choice of statistics, so we must now choose two statistics, one for each partition.

Suppose that both the partitions submit to Bolzmann's statistics. The problem now reduces itself to finding an array of n_i and m_j which renders the following functional a maximum,

$$S(N;n_i,n_2,\ldots,n_m;\ m_1,m_2,\ldots,m_p) = RN!\left(\prod_{i=1}^{m} n_i! \prod_{j=1}^{p} m_j\right)^{-1} \qquad (14)$$

under conditions

$$\sum n_i = \sum m_j = N$$

$$m \leq N; \quad p \leq N.$$

The formula (14), where R is a normalization factor gives the probability of the arrays n and m. To solve the problem of its maximization or an equivalent problem of minimization $\prod_i n_i! \prod_j m_j!$, we should take into consideration that n and m are not independent, because of (13). The technique of finding the extreme value is not discussed here (a detailed account can be found in Arapov & Šrejder 1978). Still, it is clear that the solution cannot be trivial, since the array which minimizes only one factor in (14) (this array would correspond to the so-called unit partition, see above, Sec.

2.2) maximizes another factor at the same time, so that their joint extreme value must be settled by compromise.

In crude form, the result obtained in Arapov & Šrejder (1978) can be represented as $n_a, m_a \approx \text{const}$, where n_a, m_a are coordinates of any limit element a. In other words, the maximal probability has an array which follows Zipf's Law. More accurately these results can be represented as the following inequality

$$|n_a m_a - n_b m_b| \leq \min(n_a, n_b, m_a, m_b)$$

where m_a, m_b are coordinates of any other limit element, $a \neq b$.

Consider another pair of statistics. Let Y follow Bolzmann's statistics as before, and Y*, the statistics of Example 4. In this instance, we should have to minimize

$$\prod_{i=1}^{m} n_i! \prod_{j=1}^{p} 2^{\binom{m_j}{2}} \qquad (15)$$

and the additional conditions are the same as for (14). Let us try to find an extreme value with the aid of the diagram shown in Fig. 4. Suppose the depicted array of CF minimizes (15). Let us transform the diagram as it is shown in Fig. 4a. By assumption, the depicted configuration is extreme, that is, the transformation cannot diminish the value of (15). Thus,

$$\frac{(n_i - 1)!(n_{i+1} + 1)!\, 2^{\binom{m_{j_2}-1}{2}\binom{m_{j_2}+1}{2}}}{n_i! n_{i+1}!\, 2^{\binom{m_j}{2}\binom{m_j}{2}}} \geq 1,$$

whence

$$\frac{n_{i+1}}{n_i} \geq 2^{-1} - \frac{1}{n_i} \cdot \qquad (16)$$

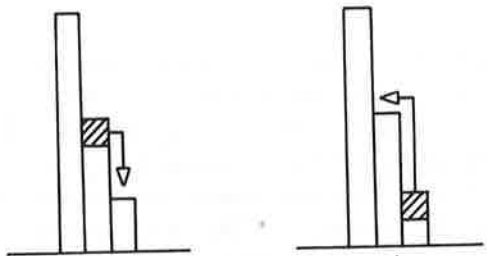


Fig. 4.

Next, applying the transformation shown in Fig. 4b and reasoning in the same way, we arrive at the following inequality,

$$\frac{(n_i + 1)!\,(n_{i+1} - 1)\; 2^{\binom{m_j}{2}}{}_2\binom{m_i}{2}}{n_i!\,n_{i+1}!\; 2^{\binom{m_j-1}{2}}{}_2\binom{m_i+1}{2}} \geq 1$$

whence

$$\frac{n_{i+1}}{n_i} \leq \left(2 - \frac{1}{n_{i+1}}\right)^{-1} \qquad (17)$$

Combining (16) and (17), we get

$$\left(2 - \frac{1}{n_{i+1}}\right)^{-1} \geq \frac{n_{i+1}}{n_i} \geq 2^{-1} - \frac{1}{n_i} \qquad (18)$$

Thus the most probable distribution of n_i is approximately a geometric progression with common ratio 1/2 (cf. Krylov & Jakubovskaja 1977). The distribution of the cell frequencies of the co-partition Y* would be logarithmic, of course.

Let us now consider the case where one partition, say Y, follows Bolzmann's statistics and another, Krylov's modification of the statistics of Example 4. In this case the most probable distribution has a more complicated character: the size of larger cells decreases approximately in geometric progression, yet for

small classes the function changes more rapidly. A similar "droop" often occurs in frequency-rank distributions of phonemes and letters (the result was discovered by Ju.K. Krylov).

Finally, if the statistics of Example 4 is chosen for both the partition Y and the co-partition Y*, then the resulting distribution is linear (an arithmetical progression).

3. Conclusion

The approach to quantitative study of language outlined here is based on the holistic concept of the closed, balanced text. For such a text "in a state of equilibrium" we can theoretically derive several conceivable laws of frequency distribution, some of them discovered empirically long ago.

Some puzzling problems concerning this approach have been left untouched in our paper. For instance, we have not dealt with the comparison of empirical data and the derived distributions. The problem can be solved partially by introducing some fitting parameters. The origin and role of such parameters as γ in Zipf's formula (5) has been investigated in our previous works (see Arapov 1977, 1978, Arapov & Šrejder 1977, 1978).

However there are situations in which by including the fitting parameters in formulas, we obscure the real problems rather than solve them. Thus with regard to the highest frequencies in Zipf's distribution, the real cause of their deviation from the Zipfian curve seems to be connected with additional restrictions on the use of auxiliary words which have high frequencies. If Zipf's Law held precisely for these words also, then the inner structure of any text (including the length of sentences, their syntactic features etc.) would be controlled by the physical size of the text. The derivation of the law cited does not take into account this restriction and any attempt to counterbalance that omission by simply introducing an additional parameter like Mandelbrot's B in (5') must be abortive (detailed argumentation is given in Arapov 1977).

Not only all limitations and restrictions imposed on the classification of the text elements, but also the principles of selecting statistics themselves have been unknown as yet. We conjecture only that the articulation of the text into parts whose inventory is limited (e.g. phonemes) involves statistics which are typified by that of Example 4, whereas the statistics of Bolzmann's type are indicated when this inventory is unbounded.

Another difficult problem which is far from solution relates to the choice of texts to which the outlined theory can be applied. Our working hypothesis is that such balanced texts are ones which have a value of the cultural event; this hypothesis is substantiated by the data collected in several works (see Orlov 1976, 1978, Arapov 1977, 1978). But what is a cultural event, or a historical event in general? The death of Charlemagne is a historical event, of course, but what about the wedding of his carpenter? The criteria of text integrity are necessarily vague and debatable. The epistemological status of text integrity is similar to that of statistical homogeneity. However, the homogeneity of an infinite population must be postulated, whereas integrity is a property of the constructive, finite object that can be studied by the inductive method.

NOTES

1. There is a small difference between the described and the original Bolzmann's statistics; unlike the former the latter admits empty cells; the same is true of Bose's statistics (see Example 3).

2. Introduced enumeration of the cells with respect to their relative size does not bear any relation to that postulated by point iv of Sec. 2.1.

REFERENCES

ARAPOV, M.V., Dve modeli rangovogo raspredelenija [Two models of the rank distribution]. Voprosy informacionnoj teorii i praktiki 4, 1977, 3-42

ARAPOV, M.V., Izmerenije leksičeskogo bogatstva tekstov [The measurement of the lexical richness of texts]. Tekst-Jezyk Poetyka. Wrocaw: Ossolineum, 1978, 185-210

ARAPOV, M.V., EFIMOVA, E.N., ŚREJDER, Ju.A., O smysle rangovych raspredelenij [On the sense of the rank distributions] NTI 1975/1, 9.20

ARAPOV, M.V., EFIMOVA, E.N., ŚREJDER, Ju.A., Rangovyje raspredelenija v tekste i jazyke [Rank distributions in text and in language]. NTI 1975/2, 3-7

ARAPOV, M.V., TER-GASPARJAN, L.I., HERZ, M.M., Sravnenije častotnych slovarej [Comparison of frequency dictionaries]. NTI 1978/4, 20-29

ARAPOV, M.V., ŚREJDER, Ju.A., Klassifikacii i rangovyje raspredelenija [Classifications and rank distributions]. NTI 1977/11-12, 15-21

ARAPOV, M.V., ŚREJDER, Ju.A., Zakon Cipfa i princip dissimetrii sistemy [Zipf's Law and the principle of dissymmetry of system]. Semiotika i informatika 10, 1978, 74-95

FRUMKINA, R.M., Statističeskije metody izučenija leksiki [Statistical methods in lexical studies]. Moskva: Nauka 1964.

HARARY, F., PALMER, E.M., Grafical Enumeration. N.-Y. - London, Academic Press 1973

KRYLOV, Ju.K., JAKUBOVSKAJA, M.D., Statističeskij analiz polisemii kak jazykovoj universalii i problema semantičeskogo toždestva slova [Statistical analysis of multiple meaning as a linguistic universal and the problem of the semantic identity of word]. NTI 1977/3, 1-6

LEVIČ, A.P., Informacia kak struktura sistem [Information as the structure of system]. Semiotika i informatika 10, 1978, 116-132

MANDELBROT, B., Adaptation d'un message à la ligne de transmission. Comptes rendus des séances hebdomadaires de l'Académie des Sciences de Paris 232, 1951, 1638-1640

ORLOV, Ju.K., Obobščennyj zakon Cipfa-Mandel'brota i častotnyje struktury informacionnych edinic različnych urovnej [The

generalized Zipf-Mandelbrot's Law and frequency structures of informational units of various levels]. In: Vyčislitel'naja lingvistika. Moskva, Nauka 1976, 179-202

ORLOV, Ju.K., Model' častotnoj structury leksiki [A model of the frequency structure of the lexicon]. Andrjuščenko V.M. (ed.) Issledovanija v oblasti vyčislitel'noj lingvistiki i lingvostatistiki. Moskva Izdatel'stvo Moskovskogo Universiteta 1978, 59-118

ŠREJDER, Ju.A., O vozmožnosti teoretičeskogo vyvoda statističeskich zakonomernostej teksta [On the possibility of deriving some statistical laws of text]. In: Problemy peredači informacii 1, Moskva Nauka 1967, 57-63

ZASORINA, L.N. (ed.), Častotnyj slovar' russkogo jazyka [Frequency dictionary of Russian]. Moskva Russkij jazyk 1977

A Theoretical Derivation of the Zipf (Pareto) Law

*Bruce M. Hill**
University of Michigan

Abstract

A simple probabilistic model is proposed for the rank-frequency and generic-specific forms of Zipf's Law. The model can be described in terms of a random number of cells with a Bose-Einstein allocation to cells. The assumptions underlying the model are discussed in several different empirical contexts.

INTRODUCTION

In this article a simple probabilistic model is proposed which leads to the rank-frequency and generic-specific forms of Zipf's Law. It is familiar in a variety of empirical contexts that a plot of $L^{(r)}$, the r^{th} largest of a set of quantities, against its rank, r, is approximately proportional to $r^{-(1+\alpha)}$, for some $\alpha>0$; and similarly, that a plot of the proportion of categories with exactly s units, say G(s)/M, is approximately proportional to $s^{-(1+\alpha)}$. These relationships will here be referred to as the rank-frequency and generic-specific forms of Zipf's Law, respectively. They seem to have been discovered independently by Willis (1) in connection with the distribution of biological genera and species, Auerbach (2) in connection with the size of cities, Pareto (3) in connection with the distribution of income, and Zipf (4) in connection with word usage frequencies. However, it was Zipf (5) who first systematically studied the phenomenon, and demonstrated its extraordinary generality. His book (5) contains a massive array of graphs drawn from a variety of empirical areas, as well as a qualitative theory of "least effort" in an attempt to explain the phenomenon.

* The author is professor of statistics at the University of Michigan, Ann Arbor, Michigan.

It is the remarkable variety and quantity of data that fit, at least approximately, such a law, that suggests the possibility of a general skeletal model applicable in a variety of scientific areas. This is not to say that very different areas need behave in the same manner, but merely that there may be some common forces operating, to a greater or lesser extent, in each. Hopefully, such a skeletal model can be fleshed out in different ways in different areas, to allow for such discrepancies as are peculiar to the individual areas where Zipf's Law seems to apply. In any case, by focusing attention upon a simple and general type of assumption, of which the phenomenon in question is a logical consequence, perhaps some insight can be gained.

The present article describes and interprets results of a model proposed by myself in (6), (7), and (8). Its purpose is to acquaint scientists interested in the phenomena, which the Zipf Law describes, with that model. Other theoretical models of interest have been proposed by Mandelbrot (9), Simon (10), and Yule (11).

MODEL

Although the same basic model yields both the rank-frequency and generic-specific forms of Zipf's Law, for concreteness, and in order to suggest different nuances, the former will be described in terms of city size, and the latter in terms of biological genera and species. There is, of course, substantial evidence that Zipf's Law holds to a good approximation in both examples (15), (11), but I do not mean to imply that the assumptions made here apply literally in either. Similarly, the language used here should be interpreted in a suggestive rather than a literal sense. With these understandings, let us begin with the model for city sizes.

Suppose that a country consists of K regions, with M_i cities and N_i people in the i^{th} such region, $i = 1,2,\ldots,K$. I shall not attempt to say precisely what is meant by such regions, since a

precise definition would of necessity be somewhat arbitrary, but have in mind geographical-political-economic entities, which although under the influence of similar underlying forces or constraints, operate to some extent autonomously. On a global scale such "regions" might in fact each be countries, while within the United States they might consist of the "North-East", "Middle-West", etc. (To illustrate the concept further, in the context of personal income, the role of a region might be played by a corporation, or even an entire industry, while in the context of word frequency usage, it might be played by a paragraph or perhaps a chapter.) The model only requires that there be some way of decomposing the country into regions for which the assumptions to be made are approximately valid.

Let $\Theta_i = M_i/N_i$ be the average number of cities, per person, in the i^{th} region. These quantities play a key role in the theory proposed here. It is assumed that the Θ_i exhibit random characteristics which can best be described in terms of a probability distribution $F(x) = Pr\{\Theta_i \leq x\}$, or more precisely, as though $\Theta_1,\ldots,\Theta_K$ were independent random variables, each with distribution F. In the applications it is the autonomy of the regions that leads to independence of the Θ_i, while the existence of similar underlying forces or constraints in the various regions, leads to the common distribution. Although such randomness of the Θ_i cannot be proved, the Θ_i are surely not deterministic, and the data are quite consistent with the above assumptions.

Next, given M_i and N_i, consider the distribution of city sizes in the i^{th} region. Let L_{ij} be the population of the j^{th} city in the i^{th} region, $j = 1,2,\ldots,M_i$, where the cities are listed in any convenient order, for example, alphabetically. Assume that the vector $(L_{i1},\ldots,L_{iM_i})$ also exhibits random characteristics, which can best be described in terms of a probability distribution for $(L_{i1},\ldots,L_{iM_i})$, conditional upon M_i and N_i. Thus N_i people are to be distributed to M_i non-empty cities according to some probabilistic mechanism. A particular such distribution, or to put it another way, a particular probabilistic allocation-scheme for allocating people to cities, plays an important role, namely, the Bose-Einstein distribution (12), familiar to physi-

cists. By this it is meant that all possible city size vectors $(L_{i1},\ldots,L_{iM_i})$ are equally likely, where the L_{ij} are positive integers and $\sum_{j=1}^{M_i} L_{ij} = N_i$. Since the total number of such vectors is $\binom{N_i-1}{M_i-1}$, it follows that

$$\Pr\{(L_{i1},\ldots,L_{iM_i}) \mid M_i,N_i\} = \binom{N_i-1}{M_i-1}^{-1}$$

for each such vector. The final assumption is that such Bose-Einstein allocations of people to cities take place independently within each region, i.e., the allocation in any one region is independent of those in other regions. The skeletal model can then be succinctly described in terms of K independent regions, the i^{th} such region having a random ratio $\Theta_i = M_i/N_i$ of cities to population, and, conditional upon M_i and N_i, a Bose-Einstein probabilistic allocation of the N_i people to the M_i cities in that region.

Under these basic assumptions, and with in addition $F(x) \sim Cx^{1/1+\alpha}$ as $x \to 0$, where C is a constant, and the symbol "~" indicates that the ratio of the two sides goes to 1, it was shown in (7) that if $L^{(1)} \geq L^{(2)} \geq \ldots \geq L^{(R)}$ are the sizes of the R largest cities in the country, and if the N_i are large, then $L^{(r)}$ will be approximately proportional to $r^{-(1+\alpha)}$, which is the rank-frequency form of Zipf's Law.

Now consider the generic-specific form of Zipf's Law. Suppose there are K families, with M_i genera and N_i species in the i^{th} such family, and let $\Theta_i = M_i/N_i$ be the average number of genera per species in that family. Let L_{ij} be the number of species in the j^{th} genus of the i^{th} family, and make the same assumptions about the Θ_i and $(L_{i1},\ldots,L_{iM_i})$ here as in the rank-frequency case for city sizes. Let $G_i(s)$ be the number of genera with exactly s species in the i^{th} family, $G(s) = \sum_{i=1}^{K} G_i(s)$, and $M = \sum_{i=1}^{K} M_i$, so that G(s)/M is the overall proportion of genera with s species. It was shown in (6) that if $F(x) \sim Cx^{\gamma}$ as $x \to 0$, then the expectation of each $G_i(s)/M_i$ is, for large N_i, approximately proportional to $s^{-(1+\gamma)}$. In (8) it was shown that G(s)/M will then be approxi-

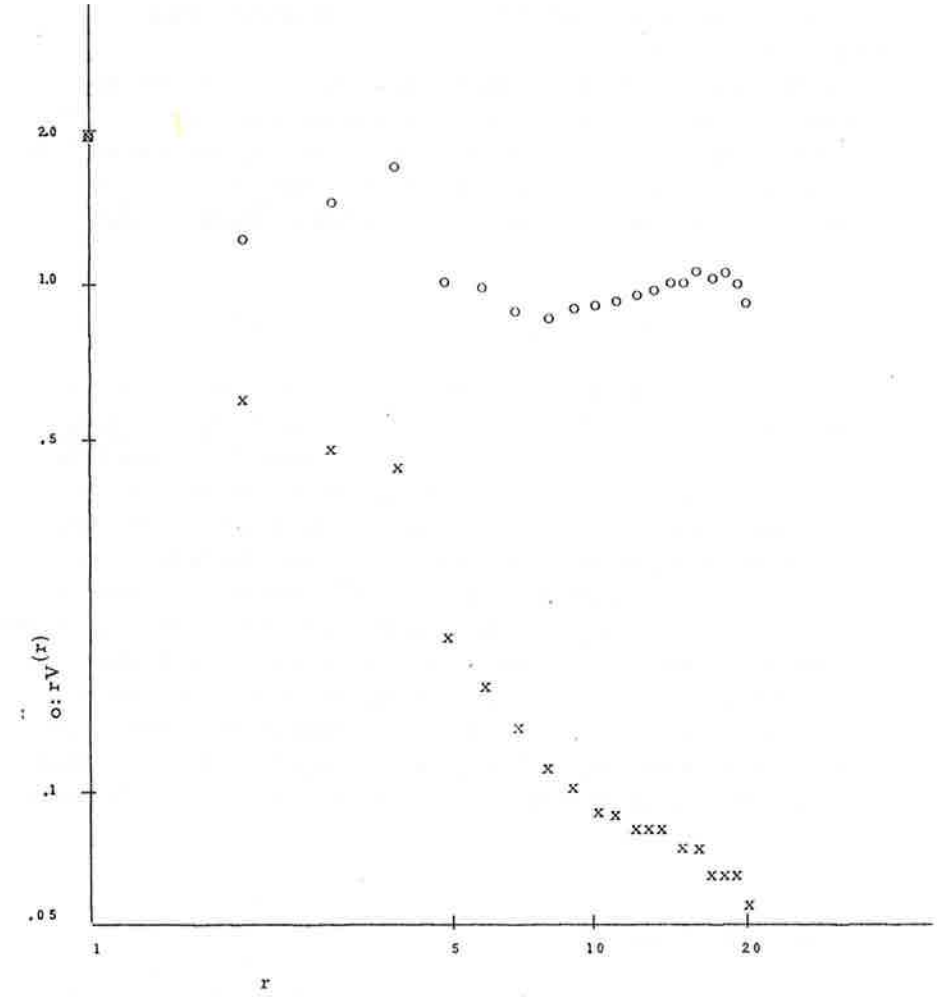


Fig. 1. Graph of $V^{(r)}$ and $rV^{(r)}$ against r.

mately proportional to $s^{-(2+\gamma)}$, which is the generic-specific form of Zipf's Law.

Particularly simple and empirically familiar results are obtained for both the rank-frequency and generic-specific forms of Zipf's Law, when $F(x) = x$, i.e., when each θ_i has the uniform distribution on the unit interval. In the rank-frequency case, if the $\ln N_i$ are nearly equal, say to $\ln(N/K)$, where $N = \sum_{i=1}^{K} N_i$, then

$$L^{(r)} \approx C[V_1+\ldots+V_r]^{-1}, \text{ for } r = 1,2,\ldots,R,$$

where the V_i are independent random variables each having the exponential density e^{-x}, and $C = K\ln(N/K)$. Since $\bar{V}_r = r^{-1}\sum_{i=1}^{r} V_i$ will be approximately 1 as soon as r is moderately large, thus $rL^{(r)}$ will be nearly constant, a result which has been noted for city sizes and a variety of other empirical phenomena (5), (4), (11). With the V_i taken from a table of random numbers, Figure 1 displays, on log-log paper, graphs of $V^{(r)}$ and $rV^{(r)}$ against r, where $V^{(r)} = [V_1+\ldots+V_r]^{-1}$. These graphs are very similar to those of Zipf and others drawn from a variety of empirical areas.

In the generic-specific case, with $F(x) = x$, then $G(s)/M$ is approximately proportional to s^{-3}. If, however, a linear regression holds between $\ln M_i$ and $\ln\theta_i$, with slope δ such that $-1 \le \delta \le 0$, then $G(s)/M$ will be approximately proportional to $s^{-(1+\alpha)}$, $0 \le \alpha \le 1$, as discussed below.

IMPLICATIONS OF THE MODEL

The mathematical implications of the above model will now be sketched. Arguments here are heuristic, intended only to supply insight into the phenomena, with rigorous statements and proofs available in the references cited.

To obtain the rank-frequency form of Zipf's Law, let $L_i^{(1)}$ be the size of the largest city in the i^{th} region, $i = 1,2,\ldots,K$. It is shown in (7) that, for large N_i,

$$\Pr\{L_i^{(1)}/\ln N_i \le y\} \approx H(y) = \int_{\theta(y)}^{1} F'(x)\,dx,$$

where $\theta(y) = 1-e^{-1/y}$. If $F(x) \sim Cx^{\gamma}$ as $x \to 0$, then $1-H(y)$ is asymptotically proportional to $y^{-\gamma}$ for large y. Now let $Z_1,\ldots,Z_K$ be any K independent random variables each having distribution H, where $1-H(y) \sim Cy^{-\gamma}$ as $y \to \infty$. If $Z^{(1)} \ge Z^{(2)} \ge \ldots \ge Z^{(K)}$ denote the ordered Z_i, then it is shown in (7) that $K(Z^{(r)})^{-\gamma}$ can be represented, approximately, as $V_1+\ldots+V_r$, where the V_i are again independent random variables having the exponential distribution. Just as was illustrated in Figure 1, for the case $F(x) = x$, it follows that a plot of $Z^{(r)}$ against r will now be approximately proportional to $r^{-\gamma}$. In particular, this will be so, with $Z_i = L_i^{(1)}/\ln N_i$, $i=1,\ldots,K$. Furthermore, if the $\ln N_i$ are nearly equal, it follows that a plot of the r^{th} largest of the $L_i^{(1)}$ against r will also be approximately proportional to $r^{-\gamma}$.

Finally, consider a plot of $L^{(r)}$ against r, $r=1,2,\ldots,R$, where $L^{(r)}$ is the size of the r^{th} largest city in the country. Clearly, if the R largest cities in the country are each the largest in some region, i.e., if the $L^{(r)}$ are a subset of the $L_i^{(1)}$, then a plot of $L^{(r)}$ against r must again follow the same Zipf Law. It is possible, of course, that the second largest city in one region is larger than sufficiently many of the $L_i^{(1)}$, so that some $L_i^{(2)}$ is among the R largest city sizes; and similarly, for some $L_i^{(3)}$, etc., where $L_i^{(r)}$ is the size of the r^{th} largest city in the i^{th} region. However, it is shown in (7) that as soon as K is moderately large, the probability of such events becomes small, so that the $L^{(r)}$ are with high probability a subset of the $L_i^{(1)}$. If $\gamma = (1+\alpha)^{-1}$, then $L^{(r)}$ will be approximately proportional to $r^{-(1+\alpha)}$, $r=1,2,\ldots,R$, as desired.

Now consider the generic-specific case. It was shown in (6) that each $G_i(s)/M_i$ is, for large N_i, approximately equal to $\theta_i(1-\theta_i)^{s-1}$, where the $\theta_i = M_i/N_i$ are, as before, independent with approximate distribution F. Hence

$$G'(s)/M = \left[\sum_{i=1}^{K} \frac{N_i}{N}\frac{M_i}{N_i}\frac{G_i(s)}{M_i}\right] \Big/ \frac{M}{N}$$

$$\approx \left[\sum_{i=1}^{K} \frac{N_i}{N} \theta_i^2 (1-\theta_i)^{s-1} \right] / \sum_{i=1}^{K} \frac{N_i}{N} \theta_i .$$

It is shown in (8) that the law of large numbers applies, so that $G(s)/M$ is approximately $[\int_0^1 x^2(1-x)^{s-1}F'(x)dx]/\int_0^1 xF'(x)dx$, if K is moderately large. For $F(x) \sim Cx^{\gamma}$ as $x \to 0$, it then follows that $G(s)/M$ will be approximately proportional to $s^{-(2+\gamma)}$. A more general result derived in (8) is that if a linear regression holds between $\ln M_i$ and $\ln \theta_i$, say

$$\ln M_i = C + \delta \ln \theta_i + E_i, \quad i = 1,\ldots,K,$$

where the E_i are independent errors, then $G(s)/M$ is approximately proportional to $s^{-(1+\delta+\gamma)}$ for $F(x)$ asymptotically proportional to x^{γ} as $x \to 0$. The case $\delta = 1$ is equivalent to the case discussed above, where the distribution of θ_i does not depend upon N_i for large N_i; while the case $\delta = 0$ implies that the distribution of θ_i does not depend upon M_i for large N_i. As shown in (8), there is some empirical evidence that $\delta < 0$, so that large values of θ_i tend to be associated with small M_i.

It is worth remarking that generally speaking the approximations to both the rank-frequency and generic-specific forms of Zipf's Law that are obtained in this theory are, if anything, somewhat better than might have been anticipated, holding for substantial chunks of the domain of the variables r and s, respectively.

DISCUSSION OF ASSUMPTIONS

I have presented a simple mathematical model, of which both the rank-frequency and generic-specific forms of Zipf's Law are a consequence. It is therefore a model which yields results that are in substantial accord with great quantities of data drawn from a variety of scientific areas. For this reason alone it would be of interest to inquire as to the plausibility of the assumptions of that model.

The assumption of approximately independent regions (families), i.e., in which the θ_i and allocation of people to cities in one region (species to genera in one family) are approximately independent of those in other regions (families), seems modest enough, although as remarked earlier there is some flexibility as to the precise definition of regions (families). Next, for those willing to allow the existence of stochastic elements, the assumption that the θ_i behave as though they were a random sample from some distribution F, would also appear to be a modest one. There is empirical evidence (8) that such a distribution might be of the form $F(x) \sim Cx^{\gamma}$ as $x \to 0$, for some $\gamma > 0$, but in any case the family of such distributions is a large one, so this assumption is not particularly restrictive. Moreover, the values $0 \leq \alpha \leq 1$, for the rank-frequency form of Zipf's Law, proportional to $r^{-(1+\alpha)}$, correspond to particularly simple distributions F, e.g., $\alpha = 0$ corresponds to $\gamma = 1$ (which includes the uniform distribution), while $\alpha = 1$ corresponds to $\gamma = \frac{1}{2}$ (which includes the arc sine distribution). Since there is substantial evidence that $0 \leq \alpha \leq 1$ in a variety of scientific areas, this seems to me to lend considerable support to the entire model. By contrast, if in order to obtain Zipf's Law it had been necessary to choose F of some unusual and exotic form, then this would have made the model somewhat less plausible. In the generic-specific case, under the regression model for $\ln M_i$ versus $\ln \theta_i$ described above, Zipf's Law again follows for plausible choices of γ and δ, e.g., $\gamma = 1$, $\delta = -\frac{1}{2}$, yields $G(s)/M$ proportional to $s^{-3/2}$, while $\gamma = 1$, $\delta = 0$, yields $G(s)/M$ proportional to s^{-2}, and there is empirical evidence (5), (11), for such behavior of $G(s)/M$.

In my opinion, the really critical assumption of the model is that of a Bose-Einstein allocation of people to cities within regions (species to genera within families). Although the Bose-Einstein distribution need not hold literally to obtain Zipf's Law, gross violations will lead to a very different result. For example, if the Bose-Einstein allocation were replaced by a Maxwell-Boltzmann allocation (i.e., if the $L_{ij}-1$ had a multinomial distribution with $N_i - M_i$ trials and probability $p_{ij} = M_i^{-1}$ for each of the M_i cells in the i^{th} region), then, as shown in (7), $L_i^{(1)}[\ln N_i / \ln\ln N_i]^{-1}$

converges in probability to 1 as N_i grows large, and a plot of $L^{(r)}$ against r will be much flatter than under Zipf's Law. The essential difference between the two allocation schemes seems to be that under the Bose-Einstein model there will be much greater variability in city sizes within a region, including typically a few extremely large cities, while under the Maxwell-Boltzmann model such city sizes tend to be more nearly equal. On the other hand, the Maxwell-Boltzmann model is somewhat anomalous. For, as shown in Hill (7, p. 1024), both the Bose-Einstein and Maxwell-Boltzmann models arise from a symmetrical Dirichlet-Multinomial model, with the Bose-Einstein corresponding to $\alpha_1 = \ldots = \alpha_M = 1$, for the parameter value, and the Maxwell-Boltzmann corresponding to $\alpha_1 = \ldots = \alpha_M$ tending to infinity. It was shown by W. Chen (13), (14), in his University of Michigan Doctoral Dissertation, that in fact, all symmetrical Dirichlet-Multinomial models give rise to a Zipf's Law, excluding only the Maxwell-Boltzmann. This demonstrates the robustness of the conclusions based upon the Bose-Einstein model within the class of symmetrical Dirichlet-Multinomial models. It appears that still other allocation schemes which yield substantial variability of city sizes within regions will also yield Zipf's Law, but this has not yet been proved. In Hill (14) actual city sizes in the United States of America are analysed using a new and simple form of statistical analysis that vividly demonstrates the surprisingly good fit of the data to the model proposed in this article.

I anticipate that many scientists will reject the Bose-Einstein model, even as an approximation, in regard to their particular area, on the grounds of its being too simple to describe the complex phenomena in question. I sympathize with such a view, but at the same time I wonder whether this is not possibly a matter of not seeing the forest because of the trees. For example, in regard to city size, to someone who is familiar with the enormous complexity involved in birth rates, migration, etc., the Bose-Einstein Law will surely seem far too simple. Similarly, to a taxonomist, familiar with all the tedious and complex analysis that goes into the classification of organisms, not to mention the complexity of the underlying evolutionary process, the Bose-Einstein distribution must seem unreasonably simple. But perhaps such intimate knowledge of the details only obscures what, from the view of an outsider, might in some respects be a relatively simple process. It does not seem too outlandish to imagine that the net result of a large number of complex interacting forces might be such a simple process. It would then be of interest to try to determine how the Bose-Einstein allocation might arise out of such complexity, and such a study is now underway. In any case, the skeletal model presented here, because of its simplicity, generality, and accord with vast quantities and varieties of data, would seem worthy of study, if only to point out what discrepancies there are from its implications, in each of the areas where Zipf's Law is applicable.

A closing remark is that ultimately one is dealing with a dynamic process, where changes occur as a function of time. For example, Zipf's Law for city sizes has held until very recently, but the development of suburbia seems to have altered matters to a certain extent. A more sophisticated model than that presented here would deal with the dynamics of the situation, and not merely the one-dimensional view obtained at a given point in time.

REFERENCES

1. WILLIS, J.C., Age and Area (Cambridge University Press, Cambridge, 1922)

2. AUERBACH, F., "Das Gesetz der Bevölkerungskonzentration," Petermans Mitteilungen 59, 74, 1913

3. PARETO, V., Cours d'Economie Politique (Lausanne and Paris, 1897)

4. ZIPF, G.K., The Psychobiology of Language (Houghton-Miflin, Boston, 1935)

5. ZIPF, G.K., Human Behavior and the Principle of Least Effort. Addison-Wesly Press, Cambridge, 1949)

6. HILL, B.M., "Zipf's Law and prior distributions for the composition of a population,". J. Amer. Stat. Assn. 65, 1220-1232, 1970

7. HILL, B.M., "The rank-frequency form of Zipf's Law,". J. Amer. Stat. Assn. 69, 1017-1026, 1974

8. HILL, B.M. and WOODROOFE, M.B., "Stronger forms of Zipf's Law," J. Amer. Stat. Assn. 70, 212-219, 1975

9. MANDELBROT, B., "The Pareto-Levy law and the distribution of income," International Economic Review 1, 79, 1960

10. SIMON, H.A., "On a class of skew distribution functions," Biometrika 42, 191, 1953

11. YULE, G.U., "A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F. R.S., Philosophical Transactions B 213, 21, 1924

12. FELLER, W., An Introduction to Probability Theory and its Applications, Vol. 1, 2nd ed. (John Wiley, New York, 1957)

13. WEN CHEN, "On Zipf's Law," Doctoral Dissertation, University of Michigan, 1978

14. WEN CHEN, "On the weak form of Zipf's Law,". Journal of Applied Probability 17, 611-622, 1980

15. HILL, B.M., "A simple general approach to inference about the tail of a distribution," Annals of Statistics 3, 1163-1174, 1975

Quantitative Analysis in the Humanities: the Advantage of Ranking Techniques

Bertram C. Brookes

1. PEOPLE AND PARTICLES

The use of the computer applied to concordances, vocabulary studies, morphology, syntactics and stylistic analysis has given rise to archival collections of complete 'machine-readable' texts which, having served their initial purposes, are made available to other scholars (Hockey 1980). There are now large collections of such texts in classical Greek, Latin, Hebrew, Old English, American English, French and doubtless many other languages. The existence of such texts invites quantitative analysis on a large scale which, without the computer, would not have been practicable.

When quantitative problems arise from such work, the linguist naturally turns first to the computer techniques of numerical analysis which, in the form of ready-to-use 'packages', are provided by most computer centres. Unfortunately, however, these statistical programs are based on the techniques of analysis developed over the past 100 years for use in the physical sciences. If the question the linguist asks happens to fit into the physicalist schema, he will find a large repertoire of programs to serve his purpose and all the expert help in applying them he may need. But there is a risk that only those questions which can be answered by physicalist techniques will then be asked. Yet the humanist or social scientist may have problems of a numerical kind which have no precedents in the physicalist repertoire and whose analysis escapes the mesh of physicalist statistics or the interest of professional statisticians.

In this paper I shall prove, and then exemplify, how physicalist techniques require the 'raw data' usually available in the social sciences to be stripped of humanistic information before the numerical analysis begins. It has therefore become necessary

for humanists themselves to begin the task of developing their own quantitative techniques because that is the only way in which their needs can be met.

Most of those well qualified in mathematics or statistics and who have become aware of the problems of quantification peculiar to the humanities have sought for techniques which are sophisticated elaborations of established physicalist techniques. Such an approach not only makes it more difficult for the humanist by making him more dependent on expert help but it also has so far failed to provide useful techniques. My own alternative approach has been to look for mathematically more primitive than more sophisticated techniques and I believe that I have been able to establish the basis of a simple calculus which could be useful only in the humanities and social sciences.

The world of the physical sciences embraces immense sets of similar particles which interact according to inexorable laws of high generality. The physical entities of which the remotest galaxies are composed are closely related to those we find on Earth. The laws of physics, it seems, have held the physical cosmos in their grasp ever since time began. The humanist, on the other hand, is concerned with the behaviour of humans, a race of organic beings which, on the time-scale of the cosmos, have emerged only recently and only, it seems, on a cosmically insignificant planet of a sun similar to billions of others visible in the night sky.

The laws of human behaviour and performance are thus less general, less rigorous and less readily accessible to observation than the laws of purely physical entities. Every human is unique, even to his fingerprints. Every human institution, every human society is unique. Every geographical location, every period of human history have their unique characteristics. And human life is always changing too. So, to the humanist, human differences are much more evident, and much more interesting, than human uniformities.

It therefore seems to be possible that techniques of quantitative analysis designed for the primary objective of capturing the uniformities of the physical world are not well suited for the

analytical exploration of the world of human thought and behaviour. People are not particles. We need new techniques.

2. THE EMPIRICAL INFORMATION DISCARDED BY PHYSICALIST TECHNIQUES

My main argument rests on an application of information theory. But before proving the theorem, I have to describe the kind of data to which it applies.

My example of linguistic data is provided by a frequency list of the Hebrew words which occur in the biblical text of Leviticus, Numbers and Deuteronomy published by Morris and James (1975) which I shall hereafter refer to as LND. The data were derived by transliterating and coding the Hebrew forms into the Roman alphabet and punching the coded text on to tape. The frequency of occurrence of each verbal form was then counted by computer. The raw data are published as a computer print-out which occupies 48 pages of the publication cited above. In Table 1 I give only the first and last 5 entries to show the basic format of the data. The Hebrew words are given here only in their transliterated forms.

Table 1. Sample of LND print-out: frequency-rank distribution (M.K. Morris and E. James)

Rank	Word	Freq.	Cum.Freq.	Token%	Cum.Freq.%	1st Occur.
1	W	5446	5446	8.955	8.955	Lev0102
2	H	5250	10696	8.633	17.589	Lev0102
3	L	3358	14054	5.522	23.111	Lev0101
4	B	2278	16332	3.746	26.857	Lev0109
5	'T	2121	18453	2.067	30.344	Lev0101
		.....	.....	.....	.....	
2367	**ENYW	1	60808	99.993	99.831	Num3207
2368	**QR'	1	60809	99.995	99.873	Num1203
2369	**QRI'	1	60810	99.997	99.916	Num0116
2370	**JKB	1	60811	99.998	99.958	Deut2830
2371	**JPK	1	60812	100.000	100.000	Deut2107

I adopted this example for illustrative use because the authors cited above have hopefully prepared the data for the analyst. The words are ranked in order of decreasing frequency of occurrence in the text. There are 2371 different words and 60812 occurrences of them. The published data also include cumulative sums and percentages together with the text location of the first occurrence of each word. Further data include alphabetical lists of words occurring in each of the three books and, of course, discussion of the linguistic problems encountered during the process of transliteration. But in this paper I focus only on basic statistical problems arising from the set of data exemplified in Table 1.

The 2371 words of LND occur with a very wide range of frequency of occurrence which is typical of data derived from coherent texts. The most frequent word occurs 5446 times, but 825 of the 2371 words listed (over one third) occur only once. Thus the distribution of frequencies is very skew.

Though the data are presented in rank form, the physicalist would prefer to base his analysis on the corresponding frequency distribution, i.e., to list the numbers of words which occur once, twice, three times and so on. As the frequency distribution is much more compact than the frequency-rank distribution, it is given in full in Table 2. This distribution is also very skew and has a long straggling tail.

For physicalist analysis the distribution of Table 2 is still very cumbersome and so most statisticians would compact it further. By grouping the frequency classes to run from 1 - 10, 11 - 20, 21 - 30, and so on, the distribution can be compacted to the form it takes in Table 3. It still has a long straggling tail. The usual procedure would be to retain the extreme values, $f_r > 120$, during computation but to group them when testing the fit of the data to any hypothetical distribution which may be considered.

Though the frequency distribution of Table 2 was derived from, and exactly reflects, the data of Table 1 expressed as a frequency-rank distribution, it is not possible to reconstruct Table 1 given only the data of Table 2. Similarly, the approximations of Table 3 exactly reflect the data of Table 2, but it is not possible to reconstruct Table 2 given only the data of Table 3. At each stage of moving from Table 1 to Table 3 some information has been discarded. This loss of information can be quantified.

Table 2. LND data; complete frequency distribution

n	0	1	2	3	4	5	6	7	8	9
f		825	364	176	125	108	64	59	54	37
1	31	34	31	28	19	25	19	11	7	19
2	12	8	12	11	11	4	4	6	4	5
3	6	5	4	6	5	4	8	-	3	5
4	2	6	6	3	5	3	10	1	2	-
5	2	4	1	3	2	4	2	-	6	6
6	-	3	1	4	5	1	-	2	1	1
7	1	5	1	1	3	1	3	2	1	-
8	1	-	3	-	3	1	3	1	-	2
9	-	3	-	1	2*	-	1	-	4	-
10	-	1	1	-	2	1	3	4	-	1
11	1	1	-	1	-	1	2	1	-	2
12	-	1	-	-	1	-	2	1	-	1

Other values: 132,146,147,152,157,162(2),163,167,170,1-2,174, 178,179,181,184,189,194,206,215,216,217,221,250,264,276,281, 304,313,338,358,374,382,383(2),389,400,403,453,470,498,553, 810,845,863,901,946,1010,1180,1188,1257,2121,2278,3358,5250, 5446. Total: 2371

Example: *This entry indicates that 2 words occurred 94 times

Table 3. LND data: compacted frequency distribution
Class-intervals: 1-10,11-20,21-30,...

n	0	1	2	3	4	5	6	7	8	9
f		1843	205	71	42	38	28	19	18	13
1	11	14	8	6	1	2	2	5	4	3
2	1	1	3	1	-	1	-	1	1	1
3	-	1	1	-	1	-	1	-	1	4
4	1	1	-	-	-	-	1	1	-	-
5	1	-	-	-	-	-	1	-	-	-

Other values: 81,85,87,91,95,101,118,119,126,213,228,336,525, 545. Total: 2371.

If N distinguishable entities are ranked, there are N! possible rankings. The data as published realizes one of these N! possibilities. The information thus gained is, by information theory, measured as

$$I(R) = \ln N! \text{ units (nepers)} \qquad (1)$$

If the N entities are now formed into the corresponding frequency distribution with frequencies $f_1, f_2, f_3, \ldots, f_r$, with $\Sigma f_r = N$, the number of such arrangements possible is

$$\frac{N!}{f_1!\, f_2!\, f_3! \ldots f_r!}$$

The information gained when one of these arrangements is realized, I(F), is similarly measured by the logarithm of this number. So we have

$$I(F) = \ln N! - \Sigma \ln f_r! \qquad (2)$$

The information lost when converting the rank to the frequency distribution is therefore

$$I(R) - I(F) = \Sigma \ln f_r! \qquad (3)$$

This quantity is positive if any one $f_r > 1$ and is zero only when all the f_r are equal to 1, in which case we again have a rank distribution.

So in converting the rank to the corresponding frequency distribution there is a loss of information. The proportional loss is

$$\Sigma \ln f_r! \;/\; \ln N! \qquad (4)$$

For the data of Tables 1 - 3 these losses can be evaluated by substituting the appropriate values in eqn. (4). In transforming Table 1 (the complete data, of course) into Table 2, the proportional loss is 59.9 or almost 60% and in transforming the data

from Table 1 to Table 3 the proportional loss is 84.6% or 11/13.

Note on factorials. The above calculations require the evaluation of the factorials N! and $f_r!$. Tables of some such numbers are available (cf. Graf, Harring 1953) and some desk calculators offer factorial functions though both tables and calculators may have limited ranges. For large factorials, Stirling's approximation can be used:-

$$\ln z! \approx (z + \tfrac{1}{2}) \ln z - z + \tfrac{1}{2} \ln(2\pi)$$

But it should be noted that the value of log z depends on the logarithmic base adopted. As this base is sometimes 10 (decimal digits), sometimes e (nepers) and sometimes 2 (bits), it is necessary to check that all the factorials are computed with the same logarithmic base. If not, they can be reduced to the same base (it does not matter which) from the relationships:-

1 decimal digit = 2.3026 nepers = 3.3219 bits.

The information lost clearly concerns the word identities. In Table 1 every word is uniquely labelled by its rank. Those words which occur with equal frequency are listed and ranked in Hebrew alphabetical order. If the ranking of the words was performed by another group of workers who adopted the same alphabetical ranking rule, the two rank distributions would be identical. But in Table 2, the only words which could be unambiguously identified are those which occur with a frequency that is unique, i.e. only 85 of the total of 2371.

A further difference between Table 1 and the frequency distributions of Tables 2 and 3 is that the ranked distribution of Table 1 begins with the Hebrew words most frequently used in the text whereas the frequency distributions begin with the words least frequently used; there are 825 words which occur only once. Thus the rank distribution gives priority and weight to the commonest words of the language, whereas the frequency distributions invert those priorities.

The main issue arising, however, is that analysis by conventional frequency distribution analysis entails a loss of empirical information even before the analysis begins. The raw data collected by the social scientist is richer in particulars than the kinds of data collected by the physical scientist and for the analysis of which conventional statistical techniques have been devised. The only advantage of adopting these techniques is that they are readily available but the loss of information their use entails cannot be compensated for. Physicalist techniques operate on social data only by first removing the rich stratum of individualistic particulars. There is no possibility that physicalist techniques of higher sophistication can recapture this loss. The individualistic data simply escapes the mesh of physicalist analysis.

It may be objected that the theorem proved above measures information in some abstract sense irrelevant to information as a subjective or semantic entity. It does certainly measure information in an abstract sense. For example, it would take longer to transmit along any transmission channel the whole of Table 1 than the whole of Table 2. It might also be argued that any loss of the initial information from Table 1 could be off-set by extending the length of the text analysed. But the theorem shows that what is lost is information of a specific kind - that relating to individual particulars. As that individualistic information is lost irretrievably, the frequency distribution would remain incorrigibly biassed however large the sample may be.

So I conclude that physicalist techniques based on frequency distributions offer analyses which can explore the data fields of the social sciences only superficially. But we have to learn to ask questions which can be answered only by reference data of the level presented by Table 1.

3. GRAPHICAL REPRESENTATION

As the raw data are often extensive (that of LND occupies 48 pages), it may be difficult to see from the numbers what kind of distribution they conform with. It is therefore worth the trouble of drawing a graph of the data before beginning detailed analysis. The graph itself may suggest hypotheses that would otherwise not have occurred to the analyst or it may show that other hypotheses are obviously not worth testing.

3.1. THE LORENZ GRAPH

A common way of graphing cumulated ranked data is to draw the Lorenz graph which is still widely used to display macro-economic data. This graph has linear scales each running form 0% to 100%; ranks are marked along the x-axis and cumulated scores along the y-axis. As the data are arranged with the highest score ranked first, the cumulative graph necessarily lies wholly above the diagonal OD (Fig.1). The graph will rise steeply at first from the origin and then ever less steeply to D.

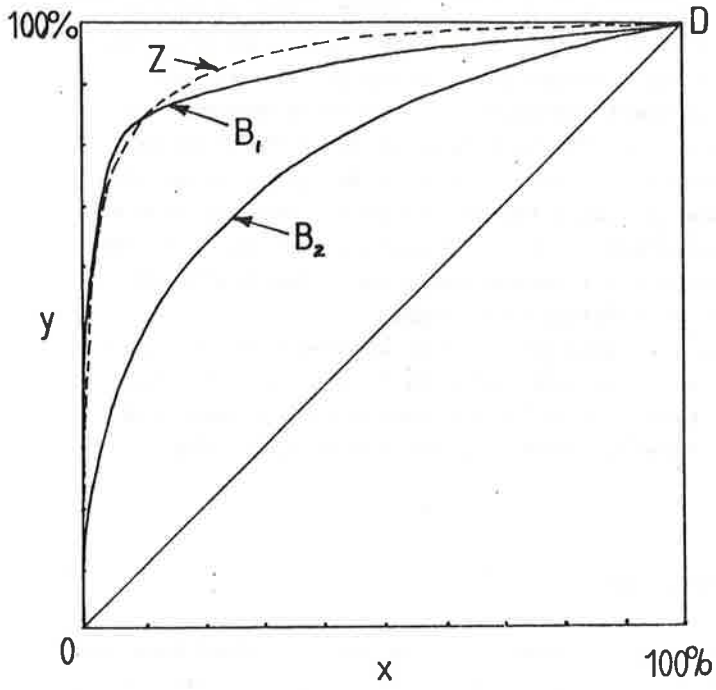


Fig. 1. The Lorenz graph has two linear scales, 0% to 100%. All ranked graphs have similar featureless forms.

From such Lorenz graphs it is possible to derive generalizations of the kind: "In a library, about 80% of the total usage of books is related to only 20% of the items most heavily used" - a rule which is often quoted. For the LND data 80% of the words of the text involve only 7.8% of the words most frequently used, as can be seen by inspection of the complete data set.

A measure of the 'concentration' of the data giving graphs of this kind was first proposed by Gini (1912). The area between the graph and the diagonal OD is compared with the area of the triangle above OD. This coefficient of concentration thus ranges in numerical value from 0 to 1. For the LND data this measure is 0.87.

But I do not recommend the use of this measure for use with social data. In the first place it is tedious to compute though it is possible to program the computation if it is needed. But, secondly and more importantly, the comparisons implicit in the measure are socially absurd, though they may have physical justifications. For example, speaking linguistically, the extreme values, 0 and 1, represent unattainable linguistic models. The text of a language with coefficient zero would be found to have exactly equal occurrences of each of its words. (This is an ideal which has importance in coding theory, however.) And the text of a language with coefficient 1 would consist of the unbroken repetition of one single word. I regard these unnatural ideals as unsuitable for work with natural languages.

There is little therefore that can be derived from the use of Lorenz graphs. One of their weaknesses is that data like those of LND always produce closely similar but featureless graphs. They are therefore not usefully discriminating nor do they suggest hypotheses to test.

3.2. THE BRADFORD GRAPH

In Fig. 2, three typical graphs of cumulative ranked data are shown. In Fig. 2(a) they are drawn as Lorenz graphs. In Fig. 2(b), the same data are plotted with both scales over the range 0 - 100%,

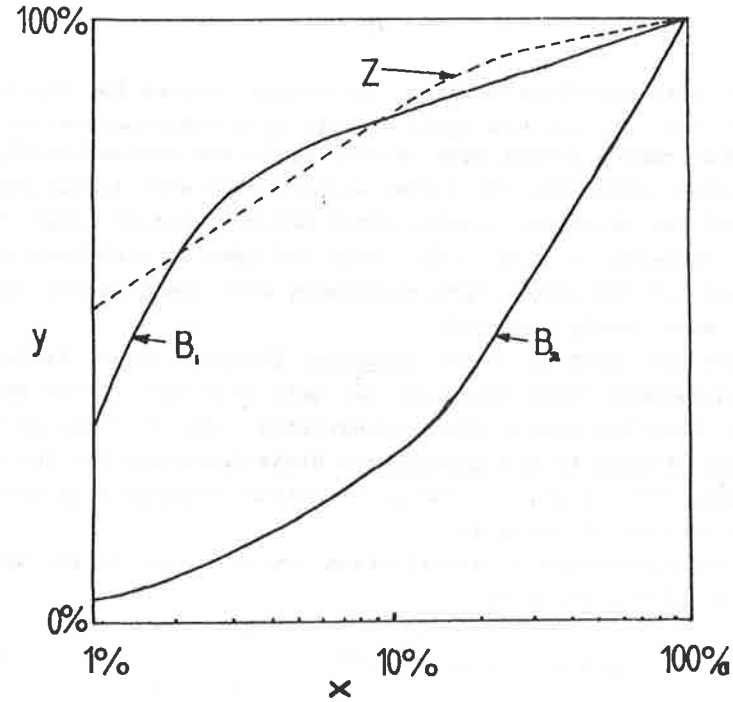


Fig. 2. The Bradford graph has one linear, one logarithmic scale. From the same data sets as Those used in the Lorenz graph, one typical Zipf and two different Bradford hybrid forms emerge.

but while the y-axis remains linear, the x-axis has a logarithmic scale.

Features which are not apparent in the Lorenz graph appear in the Bradford graph. Graph Z is that of the LND data; it shows a rising linearity which eventually falls away. Graph B_2 is that of a typical scientific bibliography; it shows a "nuclear group" of journals of high productivity of relevant papers (the rising curve) followed by a larger group of less productive journals (the linearity). Group B_1 typifies data in which the 'sources' are the natural languages used in science; here again is a nuclear group of highly productive languages followed by a long tail of much less productive languages.

There is much to be said for using a logarithmic scale for the ranks. (Suitable graph papers are available; their logarithmic scales are marked with the numbers whose logarithms are needed and which are thus provided without reference to tables.)

The main advantage of using logarithmic scales for the ranks is that the long tenuous tails of very skew distributions of the kind that ranked social data usually yield are systematically contracted. Moreover, the curves which appear when linear scales are used (as in Lorenz graphs) often become straight lines. The exact mathematical form of the graph can then be much more easily obtained and deviations from conformity with these linearities can be more easily observed.

Even when applied to the frequency distribution of Table 3, the logarithmic graph contracts the tail to a less intractable length. When the linear class-intervals 1 - 10, 11 - 20, 21 - 30, ... are changed to the logarithmic class-intervals 1 - 10, 11 - 100, 101 - 1000, ... the distribution of Table 3 is transformed to that of Table 4.

A negative binomial distribution can be fitted to the data of Table 4. It has the form

$$y = N\, h^s\, z\, (1 - kz)^{-s} \qquad (5)$$

Table 4. LND data: compacted frequency distribution
Logarithmic class-intervals: 1-10,11-100,101-1000...

n	$\log_{10} n$	f
1-10	0-1	1843
11-100	1-2	445
101-1000	2-3	74
1001-10000	3-4	9
		2371

where N = 2371, h = 0.939, k = 1 - h = 0.061 and s = 4.025. (Here h is calculated from $h = (m - 1)/v$ and $s = (m - 1)h/k$ where m is the mean and v the variance of the distribution.) The theoretical values are the coefficients of successive powers of z in the expansion of (5). In this case they yield the series 1840,552,81,... which, by the conventional χ^2-test for goodness of fit, shows that these theoretical values lie within the 5% level; the calculated value of χ^2 is 0.48 whereas the 5% and 95% limits of χ^2 for 2 degrees of freedom are 0.10 and 5.99.

However, in compacting the original LND data to those of Table 4, 90.6% of the information available for analysis has been discarded. So this fitting of the negative binomial to the original data can be regarded as only very crude. Moreover, it is not easy to relate the parameters h and s to the linguistic characteristics of Old Testament Hebrew. Yet the fact that such a physicalistic regularity has been found suggests that further regularities are worth looking for by using techniques which make fuller use of the LND data. These techniques will depend on the semi-logarithmic Bradford graph rather than on the bi-linear Lorenz graph.

4. THE EMPIRICAL LAWS OF RANKED DISTRIBUTIONS

4.1. THE SEARCH FOR THE 'ONE TRUE LAW'

Most of us have school-age memories of Boyle's (or Marriotte's) law, of Newton's law of gravitation, of Ohm's law and of Einstein's sinister $e = mc^2$. All these basic physical laws can be simply expressed and formulated. They are also exact. The development of the physical sciences still depends on the discovery of these and other quantitative laws of a similar kind.

When searching for quantitative regularities among the data provided by observations of social phenomena, we have all been conditioned, I believe, to expect to find ' the one true law' relating observable aspects of particular phenomena. But humans are not purely physical entities and, in fact, it is the non-physical manifestations of humans which engage the social sciences. Of course, we have to rely on the physical manifestations of human creativity for our observational evidence as presented by recorded language, social behaviour, the construction of human artefacts in architecture, music, the graphic arts and so on, - the man-made 'furniture' of what Karl Popper (1972) calls the world of objective knowledge.

For philosophical and other reasons I question whether any social regularities that may exist will be captured by projecting the subtle complexities of human interactions on to the epistemological plane of the physical sciences by means of conventional numerical analysis (cf. Brookes 1980a). As I have tried to demonstrate in the previous sections, conventional numerical analysis discards information which is of no account in the physical view of the world but which is of great account in the humanities and social sciences. To search for the 'one true law' of any aspect of human behaviour implies adherence to a physicalist program.

So, I argue, ranked data of the kind displayed in Table 1 offer a rich field for theoretical exploration in the social sciences. Ranking is more primitive than counting. We learn to rank, to choose the better, bigger or otherwise more attractive of two or more objects displayed to us before we learn to count "one, two, three,..." for which we need the use of language. But as ranked distributions are of no interest to the physical scientist, we find very few techniques already contrived and polished for our use. So we have to take courage and win the confidence to establish our own. Fortunately, as far as I can see, much can be done with very simple analytical instruments.

4.2. THE WORK OF SPEARMAN

The history of ranking techniques is mercifully short. They were introduced into psychometrics by C. Spearman in 1906. His need was to find a way of correlating the subjective responses of two or more humans to the same set of stimuli. His subjects might be asked, for example, to rank a set of colour samples in the order they find most attractive. If the colours are coded A,B,C,D and E, such experiments yield data of the form:-

Colour	A	B	C	D	E
S	1st	2nd	3rd	4th	5th
T	3rd	2nd	1st	5th	4th

where S and T are the two subjects asked to rank the five colours. Such rankings yield comparable sets of ordinals - 1st, 2nd, 3rd,..., which different subjects may put in different order. To find a measure of the agreement between two such sets, Spearman took the logically improper step of equating the ordinal 1st with the cardinal number 1, 2nd with 2, 3rd with 3 and so on. He thereby transmuted the sets of rank ordinals into sets of cardinal numbers. He then applied Pearson's product-moment coefficient of correlation to these numbers and thus established the coefficient of rank correlation which has the formula:-

$$\underline{r} = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad (6)$$

where n = the number of items in the ranked sets,

Σd^2 = the sum of the squared differences of corresponding rank numbers.

The Spearman coefficient is attractive because of its simplicity. But it should be noted that, apart from its highly dubious equating of ranks with numbers, it also gives equal weight to equal displacements of rank. Thus a displacement of rank from 10th to 9th is given equal weight to that of a displacement from 2nd to 1st. This is a physicalist idea which does not reflect psychological realities - as any democratic politician or football fan will aver.

The value of $\underline{r}$ is 1 when the two ranked sets are identical and is -1 when the one rank is the exact reverse of the other. For the colour-test data above we have $\Sigma\underline{d}^2 = 4+0+4+1+1 = 10$ so that with $\underline{n} = 5$ the value of $\underline{r} = 1 - 6\times10/\ 5\times24 = 0.5$.

This simplicity is so attractive that the Spearman coefficient is sometimes used on data such as that of Table 1 in which the ranks have been determined, not by subjective judgment, but by objective counting of frequencies. But to discard the frequencies is to discard valuable empirical information and is methodologically inexcusable.

An alternative measure will be proposed later in this paper.

4.3. THE WORK OF ZIPF

In 1935 the philologist G.K. Zipf (1965), who taught German at Harvard, published The psychobiology of language; an introduction to dynamic philology. He analysed statistical data on the occurrences of words in sample texts of Chinese, Latin and English and of morphemes in German. He expressed the data in two ways: first, as frequency distributions and, secondly, as frequency-rank distributions.

The frequency distributions were found to conform fairly closely with

$$w(n) = k/n^2 \qquad (7)$$

where $w(n)$ is the number of words which occur n times in the sample, while k is a constant for the sample which increases with sample size.

When such data are plotted on log-log graph paper, they conform with the equation

$$\log w(n) = \log k - 2\log n \qquad (8)$$

which is a straight line of negative slope 2.

The frequency-rank distribution conforms fairly closely with

$$f(r) = k/r \qquad (9)$$

where $f(r)$ is the frequency of occurrence of the word of rank r and k is again a constant which is determined by the sample size.

When such data are plotted on log-log paper, they conform with the equation

$$\log f(r) = \log k - \log r \qquad (10)$$

which is also a straight line but with negative slope of 1.

The advantage of plotting both of these formulations on log-log paper is that in both cases the hypothetical graph is a

straight line so that deviations from conformity with the formulation are obvious at a glance.

Zipf explored the implications of these laws of vocabulary. He also noted that many other social statistics outside linguistics conformed with these laws, at least approximately. He devoted much intellectual effort developing and propounding 'the principle of least effort' which, he argued, these laws implied. But he failed to convince others of the validity of these ideas, especially when statisticians of the period satisfied themselves that such distributions could arise from purely probabilistic considerations.

For his analytical work, Zipf adopted the frequency distribution form (7) in preference to the frequency-rank form (9). I now see this choice as a crucial mistake. Though it offered Zipf the opportunity of bringing to bear on his problems all the resources of physicalist statistics, it simultaneously opened his line of research to physicalist criticism. But he had unwittingly weakened his position by discarding the empirical information that is retained by the rank form. Zipf appreciated that his original formulation was too narrow and attempted to widen it by introducing a further parameter into the frequency formulation. But this move required him to grapple with mathematical functions of rapidly increasing complexity and this work was inconclusive. Another modification of the frequency form was proposed by Mandelbrot who at least recognised that Zipf distributions represented not 'one true law' but required division into at least two components.

The two versions of Zipf's law - the frequency and the frequency-rank forms expressed by (7) and (9) above - are not exactly equivalent. Though the frequency-rank form carries the more information, the use of this form was vitiated in the work arising from the Zipf laws by the equating of the ordinal ranks to the corresponding cardinal numbers.

4.4. THE WORK OF BRADFORD

In 1935, the year in which Zipf published his first book, another empirical law was published by S.C. Bradford who was then Director of the Science Museum Library in London. One of the services provided by the library at that time (well before computers were available) was the preparation of comprehensive bibliographies on topics in the history of science and technology. For the convenience of the authors who called for these bibliographies, the references were listed journal by journal, the journals being arranged and numbered in descending order of their productivities of relevant papers while the papers were numbered consecutively. As Director, Bradford required to see these bibliographies before they were despatched to their users. And he noted a regularity which he expressed in a rather ambiguous form (Bradford 1948) which gave rise to long debate in the journals of library science.

Bradford's difficulty, as I see it now (Brookes 1977), arose from the fact that the regularity he noted does arise in more than one form. He may have been aware of that fact but, like Zipf, he assumed that there could be only 'one true law' of bibliography and tried to accommodate all the variants he had noted in one formulation.

There is a superficial difference between Zipf's and Bradford's laws which, for some time, obscured their near-identity. As a linguist, Zipf saw no reason to cumulate his data because such sums were of no immediate linguistic interest. As a bibliographer, however, Bradford was presented with cumulative sums by the enumeration scheme he had adopted. It seems that neither Zipf nor Bradford ever became aware of the work of the other.

What Bradford noted was that when the cumulated sums of papers were plotted on a linear scale against the ranks plotted on a logarithmic scale the graph shows a rising curve which runs into a linearity. The group of the most productive journals corresponding to the rising curve he called the nucleus of the bibliography.

One advantage of the cumulated scores or frequencies first used by Bradford is that, because they help to smooth out random

variations in the individual scores, the subsequent analysis can be made more exact.

4.5. THE LOG LAW OF NUMBERS

An interesting though little,known empirical law was observed in the social use of numbers as long ago as 1906. It concerned the first digits, 1,2,3,...,9, of numbers found in social contexts such as the commercial or sports pages of our newspapers. The number zero is omitted from consideration because it plays a special role in our number system.

The law states that these digits conform with the distribution

$$p(r) = \log_{10}(1 + r) - \log_{10} r, \quad 1 \leq r \leq 9 \qquad (11)$$

The numerical values of $p(r)$ are given in Table 5. It can be seen that the larger the digit the smaller is its probability of occurrence.

Table 5. The 'anomalous' or log law of numbers

r	1	2	3	4	5	6	7	8	9
$p(r)$	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046
$P(r)$	0,301	0,477	0,602	0,699	0,778	0,845	0,903	0,954	1,000

The cumulative form of the law is

$$P(r) = \log_{10}(1 + r), \quad 1 \leq r \leq 9 \qquad (12)$$

I have argued elsewhere that there is no a priori reason why this law should not be extended beyond the first digits of the numbers observed. It is an accident of history that we have adopted the Hindu-Arab numerals with their base of 10; the Babylonians,

for example, found a base of 60 to be very convenient - a base we still acknowledge in the measurement of time. It is possible to derive the law for all numbers up to any finite integer, say b. This more general law would be expressed as

$$P(r) = \log(1 + r), \quad 1 \leq r \leq b, \tag{13}$$

where the base of the logarithms is (b + 1). (Note that logarithmic bases can be changed by using $\log_m n = 1/\log_n m$.)

This law is of theoretical interest for two reasons. First, it is the simplest of all the empirical rank laws known and therefore offers itself as a model or standard of comparison. Secondly, it is the only empirical law in which the rank positions of the entities comprising the set can be known a priori; if the set begins with 1, then the number r (in a large enough sample so that the ranks achieve stability) will be ranked rth. This distribution can therefore be used in developing a sampling theory for ranked distributions.

It is possible that the law was known to Spearman. It could have been his justification for equating ranks with their corresponding numbers. But this idea would be valid only if Spearman's ranked entities conformed with the 'log law of numbers' (as I shall call it) over the range from 1 to some number b. But this assumption is not justified.

What I have observed in social data from many sources is that some part of the distribution conforms with the law

$$P(r) = \log(a + r) - \log a \tag{14}$$

whose ranks correspond not with 1,2,3,... but with a, a + 1, a + 2, ... where the parameter a can have any positive finite value. So if we must equate ranks with numbers for the purposes of numerical analysis, it would be less offensive to put 1st = a + 1, 2nd = a + 2, 3rd = a + 3, and so on. Alternatively, if we insist on putting 1st = 1, we have to put 2nd = 1 + 1/a, 3rd = 1 + 2/a, and so on. But this step can be taken only after determining the value of a from the data.

I presume that the log law of numbers was regarded as 'anomalous' only because it had previously been assumed that the distribution would be uniform, i.e. that all the digits 1 to 9 would have an equal chance of being used in social contexts. But the log law of numbers is uniform, though over a logarithmic rather than a linear scale. The cumulative laws are thus rising linearities over segments of this logarithmic scale.

From a statistical point of view, a ranked set of data, like any other set of empirical data, must be regarded as a sample from some larger population, real or hypothetical. In different samples from the same population the same entities are likely to appear differently ranked, or at least some of them. As the sample size increases, so also the ranks of the entities become more stable. This stability begins earlier with the entities of ranks 1st, 2nd, 3rd,... But, even with independent entities like those of numbers chosen at random from a wide variety of social contexts (a process which can be simulated on the computer) surprisingly large samples are needed. The minimum sample, it has been estimated, to achieve 95% stability for the first 10 ranks is 25800 and for the first 100 ranks it is $96{,}5\times10^6$ - and numbers of this kind are independent and chosen at random from many texts.

But, even if the detail of samples varies, ranked distribution samples have closely similar overall forms from one sample to another if the entities are independent.

5. ANALYSIS OF RANKED DISTRIBUTIONS

5.1. THE STRATEGY

The log law of numbers is a theoretical ideal. I expect data to conform with this law only if the 'sources' of the counted 'items' are wholly independent and if they are all exposed to the same conditions of random selection. In fact, I have come to regard conformity with this law as a test of 'homogeneity' in social contexts.

In most cases, however, ranked data sets from social contexts cannot be presumed to be homogeneous in the above sense. The words of a coherent text, for example, are not independent nor are they random samples of the vocabulary of the language of the text.

The general strategy I advocate in analysing ranked distributions is to compare the data set with the homogeneous log law, and thus divide the set into distinguishable sub-sets. The method is best explained by the illustrative example provided by the LND data.

5.2. FITTING THE LND DATA

The graph of the LND data, Fig. 3, is typically Zipfian. It rises initially as a linearity which then declines into a curve falling away from the direction of the initial line. There is no evident point of discontinuity, merely a slow deviation from the linearity. The technique is to fit a log-linearity to the first part of the line, to extend it as far as it fits, then to fit a second line from the end of the first, extend it as far as possible, and so on, to the end of the graph.

5.3. TWO-PARAMETER FITTING

The ranked cumulated data are fitted to the equation

$$N(r) = k \ln\{(a + r)/a\} = k \ln (1 + r/a) \qquad (15)$$

which has two parameters, $\underline{k}$ and $\underline{a}$. The parameter $\underline{a}$, which marks the origin of the rank scale, is evaluated first by Wilkinson's median technique. If $N(r) = 2N(m)$, then $\underline{m}$ is the median of the ranks 1 to $\underline{r}$ and

$$k \ln\{(a + r)/a\} = 2k \ln\{(a + m)/a\} \qquad (16)$$

so that $(a + r)/a = (a + m)^2/a^2$

and $$\underline{a} = m^2/(r - 2m) \qquad (17)$$

When $\underline{r} = 100$, $\underline{N}(\underline{r}) = 43245$ (from the LND data) so $\underline{N}(\underline{m}) = 21622{,}5$. This number lies between $N(7) = 20898$ and $N(8) = 20898 + 1180$ (again from the LND data). So we have

$$\underline{m} = 7 + (21622{,}5 - 20898)/1180 = 7{,}164$$

and

$$\underline{a} = 7{,}614^2/(100 - 27{,}614) = 0{,}68 \qquad (18)$$

The value of $\underline{k}$ is determined from the fact that $N(100) = 43245$ and substituting this and the known value of $\underline{a}$ in eqn. (15) we find $\underline{k} = 8654$. The first equation is therefore

$$N(r) = 8654 \ln (1 + r/0{,}68), \quad 1 \leq r \leq 100,$$

But can the range of $\underline{r}$ be extended? The above equation is tested for values of $r < 100$, e.g., at $r = 120, 140, 160, \ldots$ If the differences between calculated and data values of $N(r)$ are sometimes positive and sometimes negative, the testing process is continued until the differences become monotonically increasing.

For the LND data, $N(195) = 49002$ (calc.) and 49003 (data) with difference $\underline{d} = -1$. But $N(200) = 49221$ (calc.) and 49216 (data) with $\underline{d} = +5$. After that point, $\underline{d}$ steadily increases; for $N(210)$ it is +12. The point $\underline{r} = 195$ was preferred to $\underline{r} = 196$ because it corresponds to the last of a run of 5 equal scores of 44. We thus have the first equation

$$N(r) = 8654 \ln\{(1 + r/0{,}68)\}, \; 1 \leq \underline{r} \leq 195 \qquad (19)$$

The end-point of this first line corresponds to only 195/2371 of the words listed in LND i.e. to 8,224% but to 49002/60812 or 80,58% of the tokens.

We now put $r_1 = r - 195$, ignore all previous data and begin

the analysis again. As before, only a modest leap forward is taken and the resulting equation is extended as far as the point at which the deviations between calculated and data values steadily increase. As this work progresses, it becomes easier because the data become smoother as r increases and so longer and more confident forward leaps can be taken.

Table 6. LND data: comparison of calculated and data values

r	S	Calc.	Data	Diff.	Error%
20	A	29552	30250	- 698	2,3
50	A	37309	37733	- 426	1,1
100	A	43429	43245	+ 4	0,01
200	B	49233	49216	+ 15	0,03
400	B	54188	54203	- 15	0,02
600	C	56497	56494	+ 3	0,01
800	C	57819	57826	- 7	0,01
1000	C	58705	58707	- 2	0,00
1250	D	59412	59395	+ 17	0,03
1500	D	59873	59895	- 22	0,04
1750	D	60215	60191	+ 24	0,04
2000	D	60488	60441	+ 47	0,08
2371	D	60812	60812	-	-

In Table 6 the results of the analysis are summarized by comparing calculated and data values at intervals over the whole range of r from 1 to 2371. It should be noted that the largest deviations occur in the first sub-set; they are large enough to justify the splitting of this first sub-set into at least two further sub-sets but I have not undertaken this further analysis.

The results are summarized in Table 7 and shown diagrammatically in Fig. 3. When using a logarithmic scale for ranks, graphs which are originally curved may sometimes be linearized by changing the rank origin, which is why the values of a were sought. The four sub-sets also appear to be more distinct than they really because both parameters, a and k, change significantly from one sub-set to the next. But as the two parameters are not wholly independent, the implications of the changes are not immediately transparent. So, in the next section, I simplify the analysis by

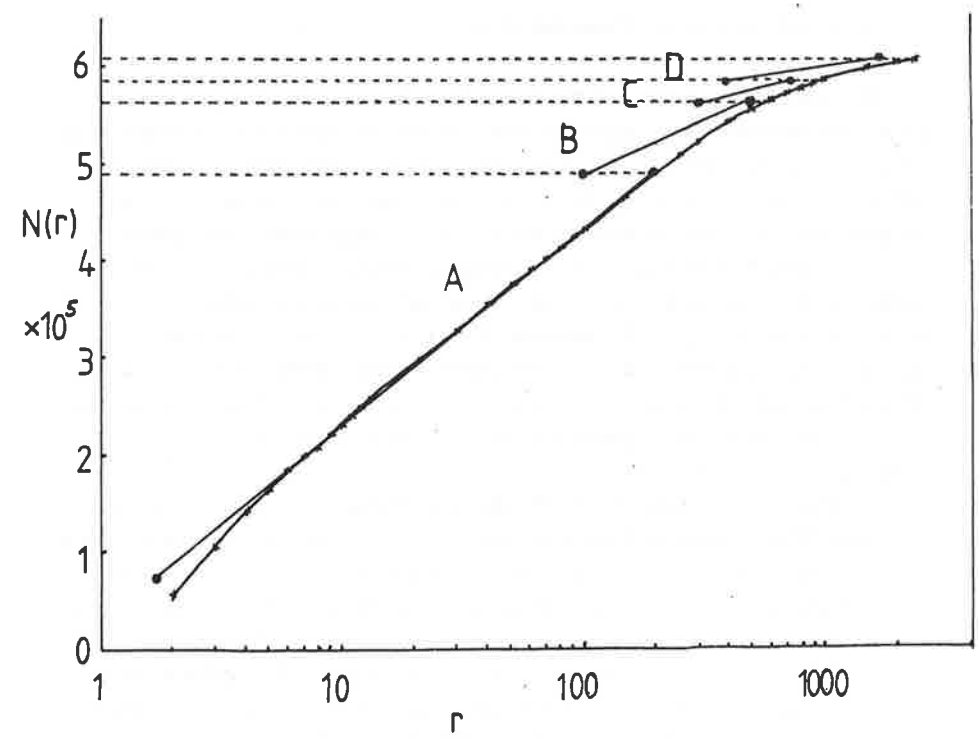


Fig. 3. Bradford graph of LND data: four linear sub-sets are shown:
A: r = 0,68 to 195,68; B: r = 196 to 497;
C: r = 305 to 720; D: r = 373 to 1733

keeping k constant and changing only a, though I also make use of the results already obtained.

Table 7. LND data: the four linearities

Sub-set	a	k	r	Σr	tokens (t)	t %	t/r
A	0,68	8654	195	195	49003	80,58	251
B	96	4537	401	596	7459	92,85	19
C	305	2659	415	1011	2284	96,60	5,5
D	373	1345	1360	2371	2066	100,00	1,5

5.4. THE GRAPH OF CONSTANT SLOPE

The second expression of eqn.(15) shows that the parameter $\underline{a}$ can be regarded not only as the origin of ranks but alternatively as a scaling factor of the ranks. In the log law of numbers, eqn.(14), the value of $\underline{a}$ is 1; in this case the ranks and the numbers on the rank scale match exactly. When ranks are marked on a logarithmic scale, the distance between successive rank marks decreases as $\underline{r}$ increases. Thus the distance between the rank marks $\underline{r}$ and $(\underline{r} + 1)$ becomes $\log(\underline{r} + 1) - \log \underline{r}$ and this distance is proportional to the contribution made by the source of rank $\underline{r}$. Equal lengths (measured in cms.) along the rank scale thus yield equal contributions to the total score and so the cumulative graph is linear.

If the graph, like that of the LND data, is not wholly linear and begins to droop before its end, this effect can be interpreted as a widening of the spaces between the ranks; as the ranks widen, their contributions to the cumulated scores declines proportionally.

The equation of the first linearity, eqn. (15), cannot be extended beyond $\underline{r}$ = 195, thus retaining its slope of 8654, except by modifying the value of $\underline{a}$ to accommodate the data.

The most direct method is retain the slope, $\underline{k}$ = 8654, of the first sub-set A for the other three sub-sets also. At the same time the rank parameters of the sub-sets are re-valued using the four linearities already obtained. The rank parameters of sub-sets B, C and D are also related to the rank origin of sub-set A. The geometry of this process is shown diagrammatically in Fig. 4 which explains how the drooping LND graph is projected on to the initial linearity.

For sub-set A, we can immediately write 195/0,68 = 286,76. This result implies that the first 195 words of the LND text are as productive as the first 286,76 numbers of the simple log law (which has a rank parameter equal to 1).

For sub-set B, we know that A and B together must satisfy the equation 8654 ln $(1 + 596/b_2)$ = 56462 where $\underline{b}_2$ is the average

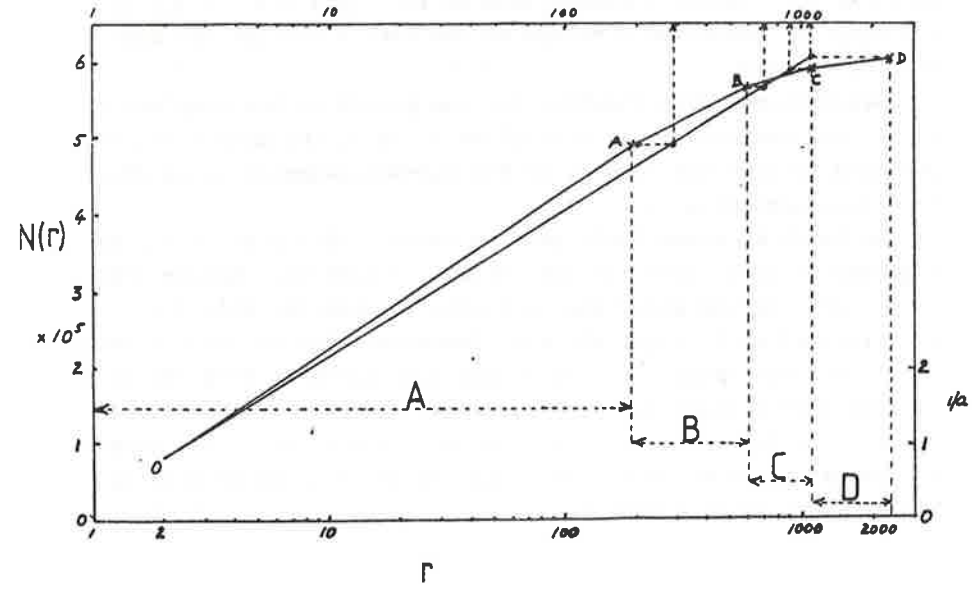


Fig. 4. Graph of LND data: the four sub-sets are projected on to the line

$$N(r) = 7825 \ln (1 + r).$$

All four rank parameters a_1 = 0,68, a_2 = 1,02, a_3 = 2,02, a_4 = 5,68 are reduced to a = 1. The A sub-set is thus extended and the other three sub-sets are contracted. Each sub-set now has a different rank density.

rank parameter over the two sub-sets. From this equation we obtain $596/\underline{b}_2$ = 680,55. But, as $\underline{b}_2$ is a weighted average of the two sub-set parameters, we need to solve the equation

$$195/0,68 + 401/a_2' = 596b_2 = 680,55$$

in which a_2' is the rank parameter for sub-set B only. We thus have $\underline{a}_2'$ = 1,02.

For the third sub-sets C and D we similarly find that $415/a_3'$ = = 205,85 and $1360/a_4'$ = 239.28 so that $\underline{a}_3'$ = 2,02 and $\underline{a}_4'$ = 5.68.

We can now write the equation of the LND graph as

$$N(r) = 8654 \ln (1 + r/a) \text{ where } \underline{a} = 0,68, \quad 1 \leq r \leq 195$$
$$= 1,02, \quad 196 \leq r \leq 596$$
$$= 2,02, \quad 597 \leq r \leq 1011$$
$$= 5,68, 1012 \leq r \leq 2371 \ldots \quad (20)$$

Thus the drooping LND graph has been rectified by projecting it on to the linearity established by the first sub-set but at the cost of re-scaling segments of the rank scale for the sub-sets B, C and D.

Alternative linearizations are now possible. The simplest is to put all the parameters a equal to 1, as in the simple log law, but then to vary the lengths of the sub-set segments along the rank axis accordingly.

As 195/0,68 = 286,76/1, the 195 sources of sub-set A can be regarded as distributed at intervals of 1/0,68 unit spaces over 286/76 unit spaces along the rank axis. Similarly, 401/1,02 = = 393,14; 415/2,02 = 205,45; and 1360/5,68 = 239,44. So the total number of unit spaces covered by all four sub-sets 1124,79. This sum provides a check on the accuracy of the analysis because we should find that 8654 ln (1 + 1124,79) = 60812, the total number of tokens in the LND text. In fact, the log product amounts to 60807, which is good enough.

5.5. THE LOG MEASURE OF CONCENTRATION

To compare the rectified linearity of the LND data, we should make the comparison with a complete simple log law rather than with the incomplete one above. The complete log law distribution has the form

$$N(r) = N \ln (1 + N)$$

where N is both the slope and the number of sources. Putting N(r) = = 60812, we have an implicit equation for N. Using a pocket calculator with logarithmic facilities, I found, by trial and error, that N = 6882. This result implies that, with only 2371 words, the LND text generates 60812 tokens whereas 6882 number sources are needed by the simple log law to generate that same number of tokens. The LND vocabulary is therefore used overall with higher concentration than the simple log law predicts. This concentration ratio can be expressed as 6882/2371 = 2,90.

The LND data-book also analyses parts of the LND text separately and so I calculated the above ratio for the separate parts

also. The results are given in Table 8. It will be seen that all except one of the ratios is greater than 1. The exception is Deut:32,33 which is also the shortest of the sections analysed and which surprised me by having a ratio of only 0,46. This result implies that its cumulative graph lies below that of the equivalent log law so that its use of vocabulary is much less concentrated than expected. Why? Is it simply because this text is the shortest? Or because its subject matter is broader? These two chapters report the last address that Moses gave to his people after their wanderings and just before he went up into the mountains overlooking the "promised land" to die. It is a powerful and wide-ranging appeal. Is that the reason? I do not know.

Table 8. LND data: concentration measures: N/W and q/m

Text	W	T	N	N/W	m	g	g/m
LND complete	2371	60812	6882	2,90	20,40	47,70	2,34
Leviticus	966	17232	2234	1,28	16,87	30,10	1,78
Numbers	1444	23275	2917	2,02	20,34	37,00	1,82
Deuteronomy	1385	20305	2584	1,87	16,75	36,23	2,16
Deut: 1-11	713	7590	1086	1,52	15,28	25,72	1,68
Deut:12-26	746	7419	1064	1,43	15,09	26,33	1,75
Deut:27-31,34	585	4278	659	1,12	16,00	23,21	1,45
Deut: 32,33	421	1018	193	0,46	46,25	19,54	0,42

The linearity analysis of Deut:32,33 is shown in Table 9 and its graphs in Fig. 5. The fact that at least one of the sample texts falls below the equivalent log law graph strengthens the case for adopting the log law as the basis of comparison. It will be seen that for the text of Deut:32,33 the three slopes, k, form an exceptional increasing sequence and that the rank spacings of its three sub-sets decrease as r increases, thus confirming its exceptionally low concentration.

The overall results of Table 8 indicate that there is a close but not exact relationship between the length of a text and the form of its graph: the longer the text, the more likely it is that its graph will be similar to that of the total LND text in

Table 9. Linearity analysis: Deut:32,33

Sub-set	a	k	r	t	Σt	R.I.*	Segment**
A	1,15	135,8	1 - 38	479	479	3.45	1,15 - 39,15
B	58,9	252	39 - 337	455	934	2,62	58,9 - 393,9
C	969	1010	338 - 421	84	1018	1,00	969 - 1053

* Rank interval compared with log law, a = 1
** Segments on the log line: homogeneous sub-sets

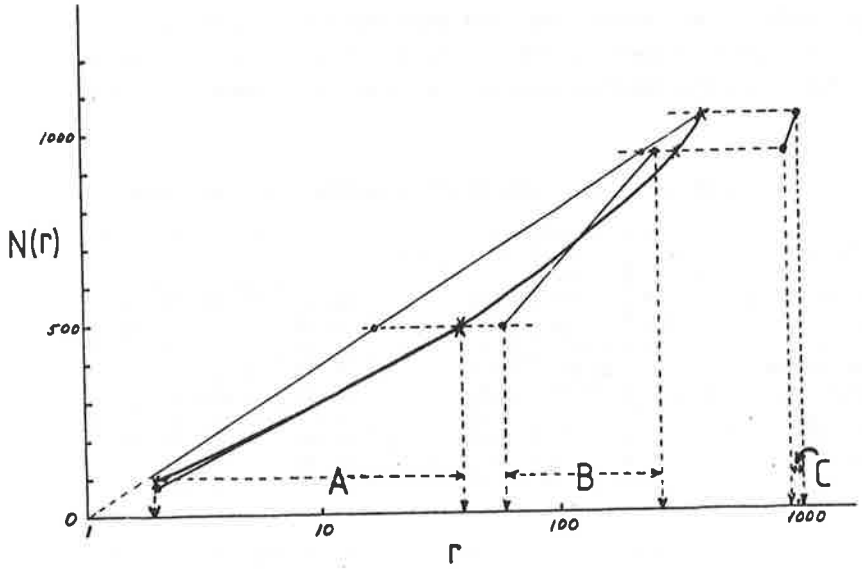


Fig. 5. Graph of the Deut:32,33 data: three linear sub-sets are identified and are projected on to the line

$$N(r) = 168{,}4 \ln (1 + r)$$

which the initial steep linearity declines into a droop. The higher will be its concentration ratio also. I believe this form to be characteristic of all coherent texts of length greater than about 1000 words and that it derives from the fact that the words of a coherent text are not statistically independent.

5.5. THE MEDIAN MEASURE OF CONCENTRATION

Cumulative graphs of ranked linguistic data, using a logarithmic scale for ranks, are usually of the form APE, shown diagrammatically in Fig. 6. The graph APE lies wholly above the straight line joining the two end-points A and E. But, sometimes, as we have seen, the graph is like AQE which lies wholly below the straight line AE.

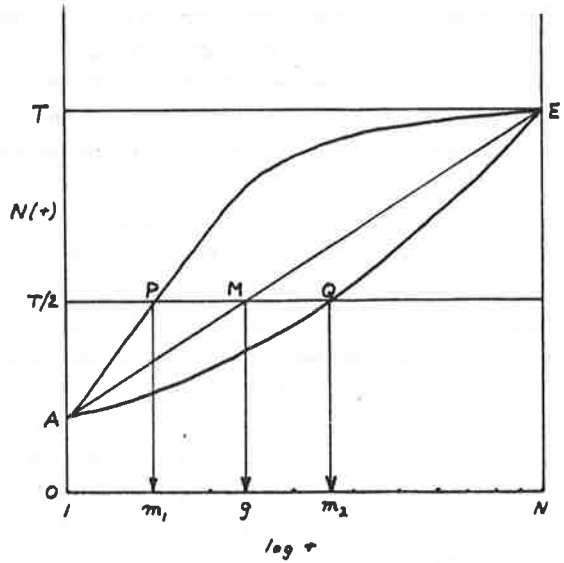


Fig. 6. The concentration coefficient g/m: the graph APE lies wholly above the line AME and $g/m_1 > 1$; the graph AQE lies wholly below the line AME and $g/m_2 < 1$.

A simple measure of concentration is obtained by comparing the median rank of the graph with the corresponding median of the linearity. If the total number of tokens is T, the median rank is that which corresponds to $\frac{1}{2}T$. This median can be found from the data directly or, less exactly, from the graph itself.

If we call m the median rank of the graph and g the median rank of the linearity, then a measure of concentration is given by g/m. This ratio is greater than 1 when the graph lies above the linearity, less than 1 when it lies below it and equal to 1 when it conforms with the simple log law.

For the LND data, the print-out shows that

$$N(21) = 50,383\%$$
$$N(20) = 49,743\%$$

so that $N(21) - N(20) = 0,640\%$. The value of $\underline{m}$ therefore lies between $\underline{r} = 21$ and $\underline{r} = 20$. By linear interpolation we have $m \doteq 21 - 0,383/0,640 = 20,40$. For the linearity we assume that $\underline{a} = 1$ so that, form eqn.(17), and substituting W for N and $\underline{g}$ for $\underline{m}$) we have $g = (W + 1)^{1/2} - 1$. For the LND data therefore, $g = 2372^{1/2} - 1 = 47,70$. The concentration of the LND data, as measured by g/m is therefore 47,70/20,40 = 2,34 which compares with 2,90 obtained by the alternative method of the previous section.

Similarly, all the other sections of the LND text yield, as before, concentrations greater than 1, except for Deut:32,33 with 0,42.

Though the matching coefficient described in the next section shows a match of 91,69% between the N/W and the g/m results in Table 8, it is advisable not to compare N/W measures directly with g/m measures.

This concentration measure, in which the graphs are compared with the diagonal in a Bradford graph, are, I suggest, more realistic and more discriminating than that of Section 3.1. based on the diagonal of a Lorenz graph.

6. THE MATCHING COEFFICIENT

6.1. THE NEED FOR A NEW MEASURE

The physicalist product-moment coefficient of correlation and the related coefficient of rank correlation both fail to capture some aspects of the comparisons of distributions of the kind provided by the LND data. We need a new measure, as simple as possible, which describes quantitatively, but also transparently,

how well one such distribution matches another. That is, of course, what the coefficients of correlation are claimed to do, but there are different kinds of matching.

The problem is analogous to that of measuring the voting behaviour in a democratic election. Psephology, the statistical study of elections, has familiarized us with the concept of voting "swing" but which I shall here call "transfer". For example, if in Elections I and II the same two candidates A and B compete for the same seat and receive votes as in the following table:

	A	B	Total
I	35320	30126	65446
II	32516	34061	66577

then A won the first but lost the second election. As compared with Election I, Election II showed a swing or transfer of votes from A to B. To measure this transfer in a standard way we need to compare equal totals. In the above example, the simplest procedure is to reduce both totals to 100 and the votes to percentages. We then get:

	A	B	Total
I	53,97	46,03	100
II	48,84	51,16	100
	+ 5,13	- 5,13	

which shows that the proportion of votes lost by A is exactly equal to the proportion gained by B, and that is 5,13%. So 5,13% is the transfer in this case. The complement of the transfer is the matching coefficient which is thus 100 - 5,13 or 94,87% for the two elections.

The transfer of votes here also led to a change of ranks. As the two ranks were exactly reversed, the coefficient of rank correlation is exactly -1. The rank correlation coefficient therefore measures the dramatic political effect of the election and the matching coefficient measures the less dramatic change of voting pattern; the net effect of the changes of votes implies that 94,87% of the votes remained unchanged.

The above example offered only two categories for comparison, but the LND data offer 2371 categories. Looking over these data, I note that the word transliterated as TM is one of 825 words which occur only once among the 60812 occurrences. As the words with equal frequancies are ranked in Hebrew alphabetical order, this word TM is uniquely ranked 2302. The word TMURA is one of 364 words which occur only twice; it is ranked 1519. A transfer of one occurrence from TMURA to TM would result in these two words exchanging their ranks, from 1519 to 2302 and from 2302 to 1519. The contribution of this exchange to the sum of the squared rank differences would be $12.783^2/2371(2371^2 - 1)$ which reduces to 1/1812. The difference made to the matching coefficient would be 1/60812. As 60812/1812 ≈ 34, the rank correlation measure is modified by this one exchange to an extent which is 34 times as great as the modification to the matching coefficient.

On the other hand, a transfer as large as (5250 - 3358)/2 = = 1892, would just fail to change the ranks of H (5250), ranked 2nd, and L(3358), ranked 3rd, but would reduce M by 1892/60812 or 3.11%.

Though such extreme effects in applying the coefficient of rank correlation are likely to cancel each other out, at least to some extent, it remains difficult to guess what any specific coefficient of rank correlation really implies. The implications are not transparent.

These uncertainties can be explained in terms of information theory. If N categories are ranked in some systematic way, as are the 2371 categories of the LND data, and we use only these ranks, as the rank correlation coefficient does, then the information produced by such ranking in ln N! = ln 2371! = 16059 nepers. If,

alternatively, we ignore the ranks, as does the matching coefficient, and depend only on the distribution of the T tokens, the information provided by any one distribution of this kind is $\ln N^T = T \ln N$ which, for the LND data, is 60812 ln 2371 = = 472574 nepers. And as 472574/16059 = 29,4 the token distribution offers much the richer information field for matching measurements. If we know the frequencies, we should make use of them.

6.2. THE LND TEXT AND THE LOG LAW OF NUMBERS

As the Hebrew text uses numbers descriptively, I noted the frequencies of occurrence of the numbers one to nine. They are given in Table 10 where they are compared with those predicted by the log law, obtained by calculating 924 $\log_{10}(r + 1)/r$ for r = = 1,2,3,...,9.

Table 10. Number of occurrences in LND: comparison with the log law

Number	1	2	3	4	5	6	7	8	9	Total
LND	253	137	90	78	121	51	179	10	5	924
Log law	278	163	115	90	73	63	54	47	43	924
Diff.	-	-	-	-	48	-	125	-	-	173
M = 1 - 173/924 = 0,8128 or 81,28%										

It can be seen that the number seven appears in LND much more feequently than the log law predicts. But this number, it is known, was regarded with special reverence by a people who at that time had little practical need for a developed numeracy in their daily lives. So seven was also sometimes used to imply many. As the frequencies gained by seven must be lost to other numbers, it is not surprising that the losses occur with eight and nine, the two numbers greater than seven. But I do not know why five occurs also very frequently.

If comparison with the log law is restricted to the first four numbers, the predicted frequencies are 240, 141, 100 and 77 respectively, for which T = 2,51% only and M = 97,49%.

The data of Table 10 are also displayed graphically in Fig. 7. It shows how T measures the misfit and M the match or fit of the two graphs as a proportion of the total area occupied by either of them.

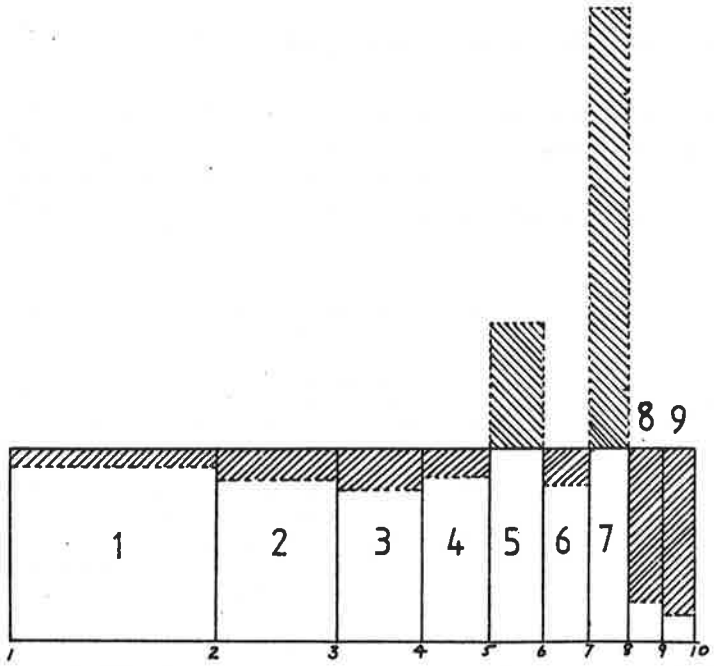


Fig. 7. The occurrence of numbers in the LND text compared with the log law of numbers using the matching coefficient. The log law predictions are represented by the columns of equal height (over the log scale) which form the large rectangle. The hatched areas represent the misfits. The proportion of the log law rectangle which remains unhatched is equal to the matching coefficient M. Here M = 81,28%

6.3. MATCHING OF LND TEXTS

The LND source-book contains data on sections of the text in which the words are arranged in Hebrew alphabetical order with their frequencies. I was thus able to compare the data of Deut: 32,33, hereafter abbreviated to D4, with the remainder of the LND text.

The complete LND text, with 2371 words and 60812 occurrences, includes the data of D4 which has only 421 words and 1018 occurrences. The LND occurrences must therefore be first reduced by subtracting the D4 occurrences. We thus compare (60812 - 1018) = = 59794 occurrences with 1018, a ratio of 58,47:1. However, the 2371 word categories can be retained for both the LND and the D4 text; it does not matter if some of the frequencies are zero.

The data pairs are then compared word by word, but only those items in which the D4 frequencies multiplied by 59 exceed the corresponding LND reduced frequencies are noted. When all these paired frequencies have been noted, the two sums, that of the D4 and that of the LND frequencies, can be found by serial addition.

The D4 items thus yielded a total of 784 and the LND items a total of 17515. We thus have

$$T = 784/1018 - 17515/59794 = 0,4772$$

so that M = 100 - 47,72 = 52,28%. This low value of M again indicates that the language of D4 is in some significant way different from that of the remainder of the LND text.

(The above calculation was "done manually" using the source-book print-outs. The calculation can be done in one of three ways: by summing those scores in which the D4 data, multiplied by 59, exceed the corresponding LND data; by summing those scores in which the LND scores exceed the corresponding D4 data, multiplied by 59; by inspecting each set of paired data, with the D4 data multiplied as before, but summing only the lesser of these two entries for each item. The first two methods involve fewer summations and evaluate T; the third method provides the evaluation

of M directly but involves 2371 items to sum. Given the LND tapes, there is no reason why this tedious but simple summation should not be mechanised.)

I thought it possible that this low value of M might arise, at least in part, because of the very different lengths of the texts being compared; one is almost 60 times as long as the other. So I next compared the D4 text with that of Deut:27-34, excluding 32,33, hereafter abbreviated to D3, which has 585 words and 4278 occurrences. The length ratio is thus reduced from 58,47 to 4728/1018 or 4,202. But here I found that T = = 3298/4278 - 257/1018 = 0,5185 so that M = 48.15%, a value which is even lower.

But I also noted that, when D3 and D4 are compared, 820 different words occur in the two texts but that only 186 of them, or 22,68% are found in both texts. This proportion seems surprisingly low.

On a simple probabilistic basis, one would expect that the 825 words of the LND text which occur only once would be more or less equally distributed over the whole text if the language were homogeneous. The number of these unique words to be expected in D4 would thus be 825 times 1018/60812, which is 14. But the number which occur in D4 is actually 85. The vocabulary of D4 is thus significantly different from that of the rest of LND.

6.4. THE MATCHING OF SAMPLES

As a check on the above results, I next compared the texts by taking samples of the words which occur in both texts. First, I chose the 10 words which occur most frequently in LND, those ranked 1st to 10th, and compared their frequencies with those of the same words in the D4 text. The calculation is given in Table 11(a) in which the Hebrew words are given as transliterated in the source-book. As the LND data include the D4 data, they must be reduced as shown in the third line of Table.

Table 11(a). Matching coefficient: LND and D4; sample

	W	H	L	B	'T	YH	'JR	+	M	KL	Total
LND	5446	5250	3358	2278	2121	1257	1188	1180	1010	946	24034
D4	100	11	40	30	3	16	4	22	32	4	262
LND*	5346	5239	3318	2248	2118	1241	1184	1158	978	942	23772
D4*	9073	998	3629	2722	272	1452	363	1994	2904	363	23772
Diff.	-	4241	-	-	1846	-	821	-	-	579	7487

LND* = LND - D4 D4* = D4×23772/262

M = 1 - 7487/23772 = 0,685 or 68,5%

Again the match is low, 68.5%, though not, of course as low as that for the case in which all the words of both texts were considered.

In Table 11(b), the LND data are similarly compared with those of D3. Here I found that M = 92,70%, a not unreasonable result.

Table 11(b). Matching coefficient: LND and D3; sample

	W	H	L	B	'T	YH	'JR	+	M	KL	Total
LND*	5029	4880	3119	2069	1969	1142	1059	1126	960	841	22194
D3	417	370	239	209	152	115	129	54	50	105	1840
D3*	5030	4463	2883	2521	1833	1387	1556	651	603	1267	22194
Diff.	-	417	236	-	136	-	-	475	357	-	1621

LND* = LND - D3 D3* = D3×22194/1840

M = 1 - 1621/22194 = 0,927 or 92,7%

When D3 and D4 are compared by means of the same 10 words, the value of M is again low - 67,12%. When 10 other words common to both D3 and D4 are used, the match remains low - 64,27%.

The only conclusion one can draw is that the language of D4 is significantly different from that of the rest of the LND text. But further elucidation of the differences calls for knowledge of Old Testament Hebrew applied to particulars.

6.5. DIFFERENCES AND LEVELS OF SIGNIFICANCE IN MATCHING

Theoretically, the range of M lies between 0% and 100%, though neither extreme is likely to arise in any practical comparison. In linguistic terms, a zero match could arise only if the two texts compared had no words at all in common. A 100% match requires both texts to use exactly the same words with exactly equal frequencies.

So most calculated values of M will lie between these extremes. In orthodox statistics several "levels of significance" have been established and are widely used. These levels are, however, purely conventional. If, for example, the difference of two means is said to be "significant at the 5% level", it means that differences equal to or exceeding that value could occur in only 5% of cases as random fluctuations in sampling the parent population, assuming statistical independence of the entities concerned. It is argued that, when a difference so large occurs, it is more reasonable to assume that the samples are not drawn from the same population and that a real difference has emerged, rather than to assume that a 1 in 20 possible random occurrence has occurred. So the convention represents a balance of probabilities.

The 5% level of significance is commonly accepted in the social sciences. Other levels - of 1%, 0,1% and so on - are used whenever the risk of error at the 5% level is unacceptable. These more discriminating levels of significance are adopted in circumstances in which, for example, an error could endanger human life.

There is no reason why similar conventions should not be adopted for use with the matching coefficient, though it would be unwise to base them probabilistic considerations. In the context of the LND data I would regard $M < 10\%$ as indicating a match; $10\% < M < 20\%$ as calling for critical comment on the difference, and $M > 20\%$ as significant, i.e. as demanding an explanation. But the level most appropriate for any specific linguistic problem must depend both on the nature of the data, on the purpose of the comparison and on the judgment of the investigator. What constitutes a "significant" difference remains, as in orthodox statistics, a matter of subjective judgment rather than of convention or of the application of a mechanical rule. The safeguard against the making of arbitrary judgments is that the grounds of the judgment be always made fully explicit so that others can examine the judgment critically.

One difference remains, however, in applying statistical analysis to physical and to humanistic problems. Whenever doubts about the outcome of a physical investigation arise, it is usually possible to repeat the statistical analysis on a further sample. In the humanities, however, the set of data we have may be all that it is possible to have. The data of LND, and of its several sub-sections, are unique. The data of Deut:32,33 cannot be replicated; they constitute a unique linguistic phenomenon. In such cases, what orthodox statisticians regard as a "sample" is also the "population". So conventions of levels of significance which are theoretically based on the assumption that large numbers of other samples can be drawn from the same population are not well-founded.

7. CONCLUDING COMMENTS

The measures I have proposed for linguistics are, I hope, both simple and transparent. Moreover, they are designed to make use of the rich empirical data that linguistics provides but much of which orthodox statistics ignores or discards.

The exception might be the use of the logarithmic scale for ranks. There are methodological reasons for adopting this idea which I have not explained in this paper, mainly because I have recently published them elsewhere (Brookes 1980). Though ranking is widely used in social affairs and is, I believe, deeply embedded in our cognitive processes, the use of ranks has been ignored in the development of statistics except in rank correlation which, I argue, is founded on the logically improper step of equating the rank ordinals with their corresponding cardinal numbers.

The present need, as I see it, is practice in the use of the new measures. With such practice new quantitative problems will emerge. But, with practice also, linguists and other social scientists interested in quantitative analysis, will, I trust, develop the confidence to establish a new calculus which will more directly serve their purposes than the repertoire of techniques imposed on the social sciences by a dominant physicalism intent on other matters.

APPENDIX

In this appendix I explain the reasons for separating the four sub-sets found within the LND data. But within the space available I cannot display the sets of empirical data needed to make the account as convincing as I would wish. I can here give only a general description of the basis of the analysis.

A1. THE LAW OF NUMBERS AS THE BASIC MODEL

The law of numbers (called by statisticians the "anomalous" law) is the simplest and most exact of the empirical logarithmic laws. Though statisticians regard it as a conventional frequency distribution because the digits themselves happen to rank themselves in their own natural order, it is also the simplest and most exact frequency-rank distribution. It therefore provides a model to which other rank distributions can be compared. It also lends itself very readily to computer simulation so that hypotheses can be quickly and efficiently tested.

Fig. A1 shows the theoretical graph for the digits 1 to 9. When the cumulated 'scores' are plotted at the upper ends of the corresponding rank intervals, the graph obtained is exactly linear - a perfect Zipf graph with no drooping tail - as shown by

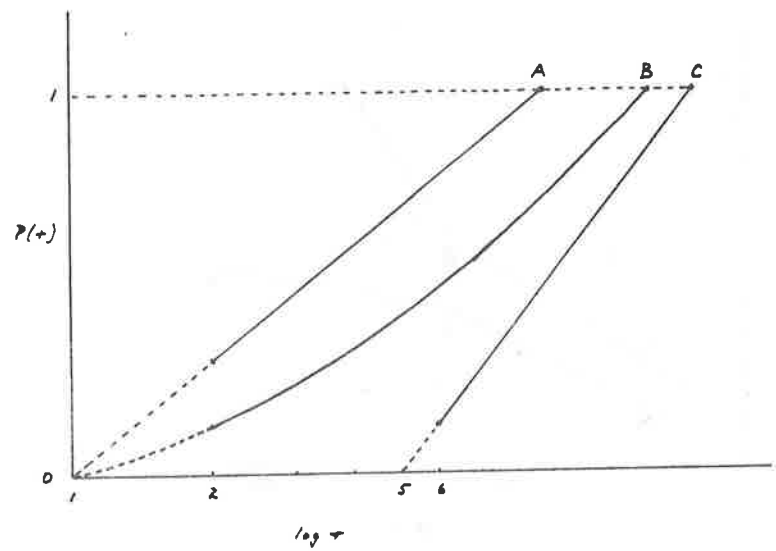


Fig. A1: Linearity and the rank origin

graph A. The theoretical graph for the numbers 5 to 20 is shown in Graph B when the origin of ranks is placed at log1: it shows a rising curve. If, however, the cumulative scores are plotted with the origin of ranks at log 5, the graph C shows that it becomes linear again.

With long experience of plotting graphs of this kind of sets of data drawn from diverse social contexts, I gradually found myself assuming that all the members of the set of sources were exposed to the same conditions of selection or performance only when linear graphs, some requiring change of rank origin, could be derived. The actual scores, of course, could be widely different but, when they were ranked and cumulatively summed, an exact log-linearity seemed to be the natural outcome.

If, for example, a set of ranked data yielded graphs such as ABC or DEF in Fig. A2, in which a marked discontinuity appeared, it would immediately suggest that such sets were hybrid or composite, i.e. that for some reason the sources corresponding to AB (or DE) were subject to different conditions from those ap-

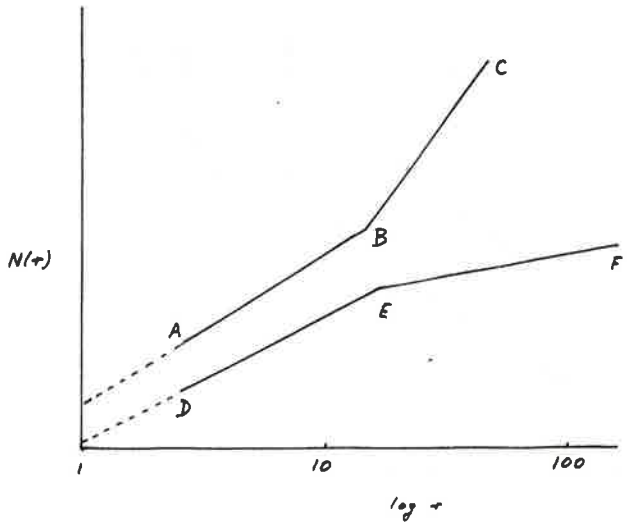


Fig. A2: Binary composite forms

plying to those of BC (or EF). Possible reasons for such differences could then be sought by closer examination of the circumstances under which the sources were acting.

A2. COMPOSITE FORMS OF THE BRADFORD LAW

The graph Bradford first observed and described had the form shown in Fig. A3. It shows a rising curve which, at C, runs into the linearity CB. Bradford noted the discontinuity and called the sources corresponding to AC the nuclear journals.

The graph is clearly composite. Two separate formulations are needed to describe it mathematically - one for the curve AC and one for the linearity CB. After some years I noted that the curve AC could also be reduced to linearity by a suitable change of the rank origin. When the origin is thus changed, the graph takes

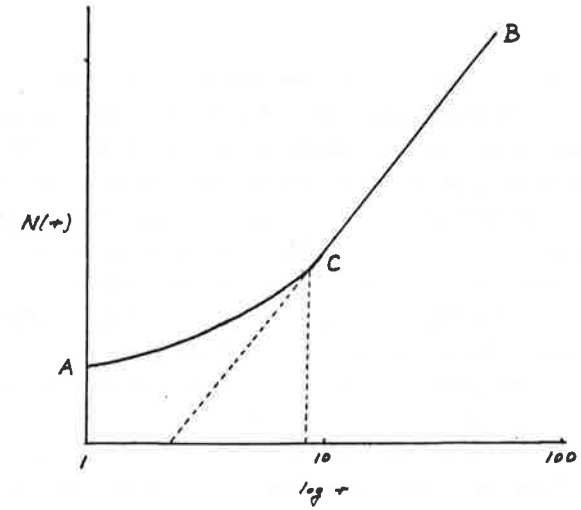


Fig. A3: The Bradford nucleus

the form of that in Fig. A4. The first part of the graph is then linear but the second part becomes a curve. The discontinuity cannot be eliminated by changes of rank origin. The graph has two components.

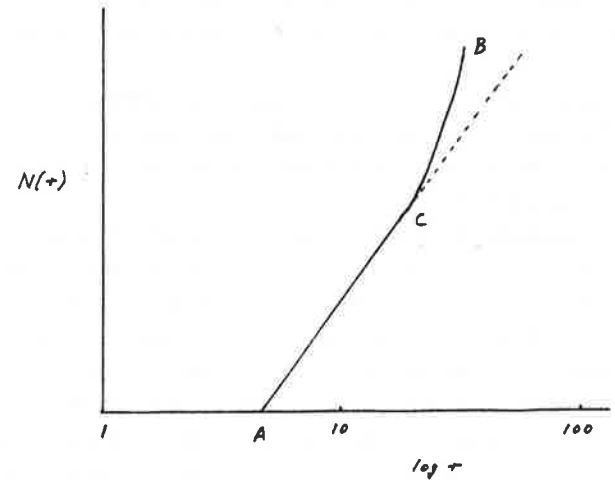


Fig. A4: The Bradford law; change of rank origin

If the bibliography concerns some scientific topic, say T, then the nuclear journals comprise those which specialise in T. They are the most prestigious journals in the field of T. They compete for the best papers on T. Their editors and referees are the most knowledgeable and the most critical in the subject. The remaining more peripheral journals may also be prestigious in other fields but they happen to be less centrally interested in the topic T. It has been somewhat cynically suggested that almost any paper on T can be published somewhere. All the author needs to do is to prepare the ranked Bradford data from his bibliography on the subject and then, if need be, to submit his paper in turn to the journals ranked 1,2,3,... on his list. As the rank of the journal thus increases, so the chance of rejection decreases. So the author finds the most prestigious journal which will publish his paper. Thus the peripheral journals publish, at least in part, papers rejected by the nuclear journals. But the reverse process, acceptance by the nuclear journals of papers rejected by the peripheral journals, is very unlikely. So, though the nuclear journals are the most productive in the field of T, there is nevertheless a net loss of papers from the nuclear to the peripheral journals. This loss or transfer can be measured by comparing the composite Bradford graph with the equivalent simple log-linearity.

Thus, in bibliographies, the hybrid form indicates that the journals can be separated into two sub-sets, the nuclear and the peripheral, and that these two sub-sets react differently in selecting the papers they publish.

A second composite form also occurs in bibliography. If the papers are classified by the natural language in which they are published rather than by the journals, so that the languages now become the 'sources', and these sources are ranked in order of decreasing frequency, the resulting graph will be of the form shown in Fig. A5. This graph rises steeply to a nuclear point and thereafter much less steeply. The nuclear languages will include English, French, German, Russian ... though not always in the same order; the peripheral languages will include, for example,

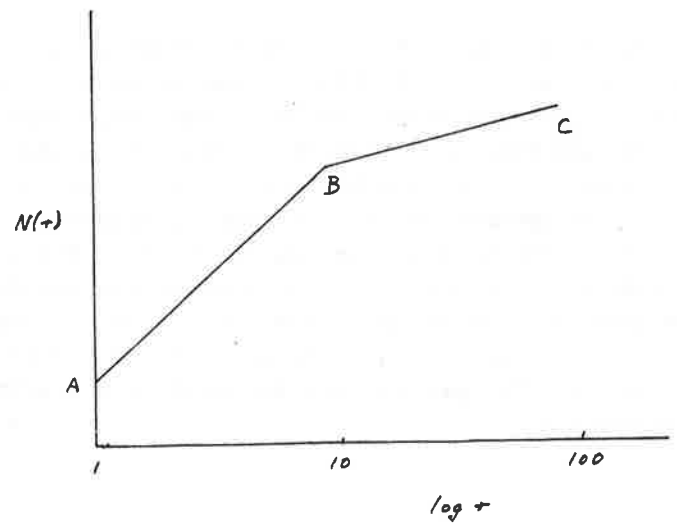


Fig. A5: The natural languages of science; a binary form

Dutch and Hungarian. But why do we not get a Zipf linearity? A Dutchman, for example, may feel that would find a wider readership for his scientific paper if he were to write in English, German or French rather than in his native tongue; a Hungarian might similarly opt for Russian. But it is highly unlikely that a scientist whose native tongue is one of the nuclear languages of science would publish his paper in one of the peripheral languages.

So again the set of sources is divisible into two sub-sets. But in this case there is a net gain of papers by the nuclear from the peripheral languages. And again this net transfer can be measured by comparing the composite graph with the equivalent simple log-linearity.

A composite form has also been noticed in conference behaviour. At scientific conferences, papers are formally read and are then discussed in open session. Though every member present is equally free to speak, members do not speak equally freely. Inevitably there are some who are a little too ready to speak, some who are

a little too reluctant. One of the chairman's duties is to elicit as wide a discussion - to ensure as many different views - as is possible. So, when he notes that two members are simultaneously expressing their wish to speak and that one of them has already contributed and the other not, he will usually give priority to the new speaker. The graph displaying conference contributions with members as the sources is like that of Fig. A2 - the nuclear point dividing the ready speakers from the shy. In this case there is a net transfer from the ready to the shy.

None of these effects are detectable by conventional frequency distribution techniques but they are graphically displayed by ranking techniques.

A3 MORE COMPLEX COMPOSITES

The composite forms discussed above have all been binary but more complex forms occur also. A musicologist recently provided me with data summarizing new recordings of classical music published over recent years. A complex type of Zipf graph emerged from the ranked data as shown in Fig. A6. Analysis similar to that applied to the LND data split the 152 recorded composers into 5 sub-groups of 3,27,51,43 and 25 members with mean recordings of 278,61,12.4,4.0 and 1.7 respectively. When the graph was compared with that of the equivalent simple log law, it indicated that the more popular composers were "over-produced" at the cost of the less well-known composers. The various interactions between the sub-sets could be quantified.

But linguistic data present new problems. Unlike the numbers drawn at random from many different sources, unlike the journals or the recordings of the classical composers, words derived from a coherent text are not wholly independent. And can whatever may be counted - words or morphemes or other linguistic forms - always be indisputably defined so that a coherent text is exhaustively analysed in precisely those forms? Such uncertainties give rise to a 'fuzziness' in linguistic data which, as I found in the

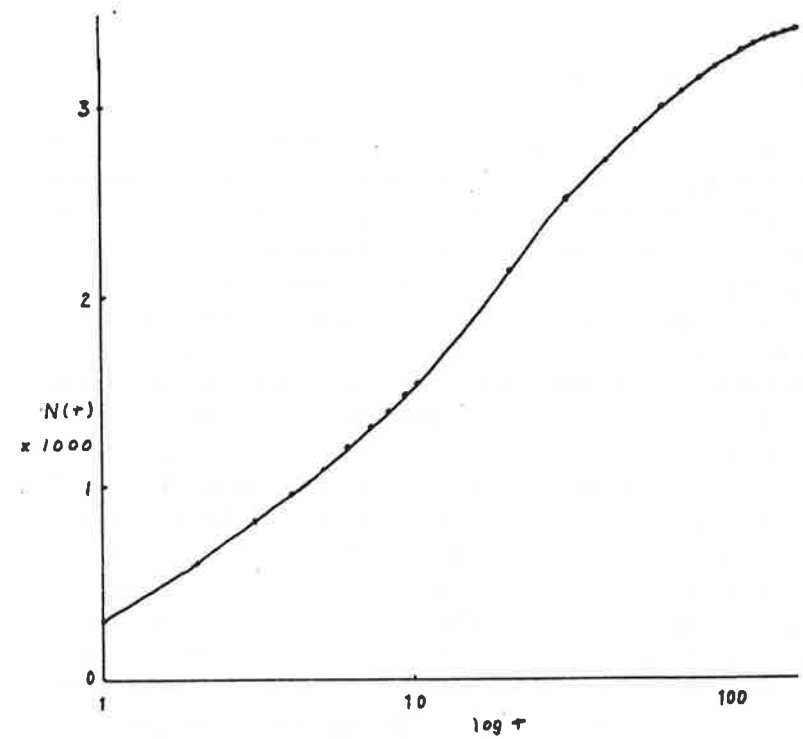


Fig. A6: Composer recordings: a 5-component form

LND analysis, tends to make the points of demarcation between successive sub-sets less sharp than those found in data from non-linguistic sources.

I much regretted that I did not know Old Testament Hebrew well enough to examine the sub-sets critically in the hope of discovering what characteristic differences they display. In what sense, for example, do the words of sub-set A, the most frequently occurring words of ranks 1 to 195, constitute the 'nuclear' vocabulary of the language? What can be said with this very limited vocabulary? What can not be said? These questions may be unanswerable, but I could not even attempt to answer them.

A4 THE LOG LAW ITSELF

The empirical laws of frequency-rank distributions have been discovered, forgotten and re-discovered several times during the present century. Each discovery has aroused a temporary interest which has waxed enthusiastically at first and has then waned as new inexplicable variants have emerged. It has long been felt, however, that some simple logarithmic regularity underlies the social manifestations in which its variants appear. There is no single all-embracing formulation which can capture all the known variants. But all the variants are reducible to composite subsets of the same basic form.

There remains the need to explain how it is that the logarithmic form is so pervasive in such contexts. The mathematical concept of the logarithm is a relatively sophisticated outcome of the pure mathematics of the 18th century applied to an idea of the early 17th century concerned with the simplification of numerical calculation. It has no primitive roots in earlier mathematics. Nor did Piaget, in his penetrating work on the development of quantitative ideas in very young children, find any trace of it - but did he look for it? One feels a need for a naturalistic explanation of a function intimately related to ranking, rather than to counting or measuring, and which is pervasive in human affairs. A possible explanation is discussed in Brookes (1980).

REFERENCES

BRADFORD, S.C., Documentation. Crosby Lockwood, London, 1948

BROOKES, B.C., Theory of the Bradford law. Journal of Documentation, 33, 1977, 180-209

BROOKES, B.C., Measurement in information science; objective and subjective space. Journal of the American Society for Information Science, 31, 248, 1980a

BROOKES, B.C., The foundations of information science. Parts I - IV. Journal of Information Science
Part I: Philosophical aspects Journal of Information Science 2(1980) 125-133
Part II: Quantitative aspects: classes of things and the challenge of human individuality 2(1980) 209-221
Part III: Quantitative aspects: objective maps and subjective landscapes 2(1980) 269-275
Part IV: Information science: the changing paradigm 3(1981) 3-12

GRAF, U. and HENNING, H.-J., Formeln und Tabellen der mathematischen Statistik. Springer-Verlag, Berlin 1953

HOCKEY, S., A guide to computer applications in the humanities. Duckworth, London, 1980

MORRIS, P.M.K. and JAMES, E., A critical word-book of Leviticus, Numbers and Deuteronomy. The Computer Bible Vo. VIII, Scholars Press, University of Montana, U.S.A. 1975

POPPER, K.R., Objective knowledge; an evolutionary approach. Clarendon Press, Oxford, 1972

WILKINSON, E.M., The ambiguity of Bradford's law. Journal of Documentation, 28, 1972, 122-130

ZIPF, G.K., The psychobiology of language; an introduction to dynamic philology. M.I.T. Press, Cambridge, Mass., U.S.A. 1965

DYNAMIK DER HÄUFIGKEITSSTRUKTUREN

Ju. K. Orlov

Der Begriff der "Rangverteilung" ist auf dem Weg, zu einem allgemein akzeptierten Begriff zu werden. Er bezieht sich auf eine bestimmte Konstanz der Form einer geordneten Menge von Häufigkeiten der Elemente in einer gegebenen Stichprobenklasse. Eine vollständige Übersicht der gegenwärtigen Vorstellungen über diesen Begriff findet man in Arapov & Efimova & Šrejder (1975).

Der Begriff der "Rangverteilung" ruft aber aus wenigstens zwei Gründen Einwände hervor. Erstens, das Wort "Verteilung" erinnert an die übliche statistische Vorstellung, für die die Sätze über Konvergenz, über statistische Stabilität usw. gelten, während in der Wirklichkeit mit der Vergrößerung des Stichprobenumfangs keine "Verbesserung" der beobachteten Häufigkeitssequenz und keine Konvergenz gegen eine "ideale" Form zustandekommt (vgl. Arapov & Efimova & Šrejder, 1975: 13). Zweitens hängt das Wort "Rang" nur mit einer gewissen Darstellungsweise für Relationen in einer Zahlenmenge zusammen. Im Grunde genommen ist der "Rang" eine Hilfsgröße ohne organischen Zusammenhang mit der Zahlenmenge. Außerdem ruft dieses Wort unerwünschte Assoziationen mit der Rangordnungsstatistik, mit Rangordnungskriterien usw. hervor. Aus diesen Gründen wird in dieser Arbeit der Begriff der "Häufigkeitsstruktur" verwendet, der die Mängel des Begriffs der Rangverteilung nicht aufweist.

Im ersten Teil der Arbeit erklären wir die stetige Approximation von diskreten Häufigkeitsstrukturen, bei der der Begriff des Ranges an sich entbehrlich ist und die es ermöglicht, die Zusammenhänge zwischen den Parametern der Struktur leicht zu finden. Diese Methode eignet sich im Grunde für die Beschreibung beliebiger Häufigkeitsstrukturen. Gleichzeitig zeigen wir den Übergang zur traditionellen Vorstellung der Häufigkeitsstruktur mit Hilfe einer rangierten Sequenz.

Im zweiten Teil werden mit dem eingeführten Formalismus die Häufigkeitsstrukturen des Zipf-Mandelbrotschen Typs analysiert.

Es wird eine allgemeine Form der Beschreibung ähnlicher Strukturen aufgestellt, ferner werden die Beziehungen zwischen den Parametern der Mandelbrotschen Formel

$$p_i = \frac{K}{(B+i)^\gamma}; \qquad i = 1,2,..v \tag{1}$$

sowie zwischen anderen beobachteten Parametern der Stichprobe, deren Häufigkeitsstruktur der Beziehung (1) folgt, dargestellt.

Im dritten Teil wird der interessante und theoretisch sowie praktisch äußerst wichtige Fall der kleinen Stichproben, zu denen beispielsweise alle lexikalischen Stichproben gehören, analysiert. Für diese Stichproben ist charakteristisch, daß ihr Vokabular das der Grundgesamtheit, aus der sie stammen, nur zu einem geringen Teil ausschöpft (als die erste Approximation an eine solche Grundgesamtheit kann man die Sprache als ganze oder eine Teilsprache betrachten). Solange wir nur über dürftige Kenntnisse der Grundgesamtheit verfügen, muß das obligatorische Vorkommen von hapax legomena (Wörter, die nur einmal vorkommen) als charakteristisches Attribut solcher Stichproben betrachtet werden. Es sind nämlich eben die hapax legomena, die die Zugehörigkeit dieser Stichproben zur Klasse der kleinen Stichproben bestimmen.

Im vierten Teil wird gezeigt, daß die Häufigkeitsstruktur kleiner Stichproben dynamisch ist und daß die Veränderungen dieser Struktur mit den Veränderungen ihres Umfangs gesetzmäßig verbunden sind. Die Dynamik der Häufigkeitsstrukturen, allgemein untersucht von Kalinin (1965), wurde bis jetzt beim Vergleich empirischer Daten mit theoretischen Modellen überhaupt nicht in Betracht gezogen. Die bekannten "Abweichungen" im Bereich der niedrigen Häufigkeiten führten zu dem Schluß, daß das Zipfsche Gesetz nur eine grobe Annäherung sei (vgl. Frumkina 1961), und stimulierten zur Ableitung komplizierter Formeln, die eine "bessere Approximation" an die empirischen Häufigkeitskurven darstellen sollten. Die im vierten Teil abgeleiteten Beziehungen beschreiben die Häufigkeitsstruktur eines statistisch homogenen Textes, der das Zipf-Mandelbrotsche Gesetz bei eindeutig bestimmtem Text-

umfang erfüllt. Es wird gezeigt, daß bei anderen Umfängen "Abweichungen an den Schweifen" üblicherweise unvermeidlich sind.

Im fünften, letzten Teil bringen wir ein Beispiel zur Berechnung des Vokabularwachstums und der Zahl der hapax legomena, sowie einen Vergleich mit empirischen Daten und mit Kalinins (1965) Berechnungen.

Die Problematik wird am linguistischen Material dargelegt, aber das konstruierte allgemeine Modell der Häufigkeitsstruktur des Zipfschen Typs kann auch bei der Lösung solcher Probleme der Informationsverarbeitungssysteme, der automatischen Steuerungssysteme, der Ökonomie, der Soziologie, der Biologie usw., die auf "Rangverteilungen" des Zipfschen Typs führen, Verwendung finden. Weitere Entwicklungen der dargelegten Methode ermöglichen es, auch zu grundsätzlich nicht-zipfschen Strukturen überzugehen, d.h. zu solchen, die nicht aus dem Zipf-Mandelbrotschen Gesetz abgeleitet werden (analog den Strukturen im vierten Teil der vorliegenden Arbeit).

1

Wir werden annehmen, daß die Häufigkeitsstruktur (≡ statistische Struktur) einer Stichprobe (Grundgesamtheit) durch die gesamte Menge der Häufigkeiten (Wahrscheinlichkeiten) der sie konstituierenden Elemente gegeben ist, ohne jegliche Anordnung der Elemente und ohne die Angabe, welche Häufigkeit zu welchem Element gehört. Diese Auffassung der Häufigkeitsstruktur unterscheidet sich wesentlich von der einer diskreten Verteilung. Sie ist in dem Sinne allgemeiner, daß unterschiedliche Verteilungen dieselbe Struktur haben können, während das Umgekehrte nicht gilt (stellt man beispielsweise im Frequenzwörterbuch die Wörter in Bezug zu den Häufigkeiten willkürlich um, so bekommt man eine andere Verteilung, während die Häufigkeitsstruktur erhalten bleibt).

Es interessierten uns im Grunde einige Beziehungen in einer Menge von Zahlen, deren Summe Eins ergibt. Diese Zahlen werden

wir bequemlichkeitshalber im weiteren als Häufigkeiten bezeichnen. Wir nehmen an, daß die Häufigkeitsstruktur einer Stichprobe bekannt ist, wenn man für ein beliebiges Intervall zeigen kann, mit welcher Wahrscheinlichkeit ein Element, dessen Häufigkeit in dieses Intervall fällt, zufällig gewählt werden kann. Um diese Aufgabe zu lösen, muß man eine stetige Approximation an die Häufigkeitsstruktur in den Fällen machen, wo die Zahl der Häufigkeitswerte groß ist und hinreichend dicht das ganze Intervall der Häufigkeitswerte der gegebenen Stichprobe ausfüllt.

Gegeben sei eine Stichprobe mit dem Umfang N aus v untereinander unterschiedlichen Elementen mit den Häufigkeiten $p_i^N = \frac{m_i}{N}$, wobei m_i die absolute Häufigkeit des i-ten Elements für i = 1,2,... ...v ist. Um weitere Überlegungen und auch den Übergang zu traditionellen Arten der Analyse der Häufigkeitsstrukturen zu ermöglichen, setzten wir voraus, daß die Menge $\{p_i^N\}$ abnehmend geordnet ist. Es interessiert uns nur das Problem der Approximation von $\{p_i^N\}$, sowie einige andere Charakteristika der Stichprobe unter der Bedingung, daß N und v so groß sind, daß die Summe der Menge der Häufigkeiten, die auf kleine Teilintervalle innerhalb des Intervalls $[p_v^N, p_1^N]$ entfallen, klein ist.

Sei $v_m(N)$ die Anzahl unterschiedlicher Elemente, von denen jedes die absolute Häufigkeit m besitzt. Entsprechend unserer Forderung bedeutet dies, daß die Größe

$$\sum_{p<\frac{m}{N}\leq p+\Delta p} v_m(N)\frac{m}{N} \tag{2}$$

bei kleinem Δp klein ist.

Wir führen folgende Funktionen ein:

$$v^N(p) = \sum_{\frac{m}{N}\leq p} v_m(N) \tag{3}$$

$$F^N(p) = \sum_{\frac{m}{N}\leq p} v_m(N)\frac{m}{N}. \tag{4}$$

Bei Anwendungen auf lexikalische Stichproben stellt Funktion (3) die Zahl unterschiedlicher Wörter, deren relative Häufigkeiten in der gegebenen Stichprobe p nicht übersteigen, und Funktion (4) die Wahrscheinlichkeit der zufälligen Wahl eines beliebigen Wortes aus dieser Stichprobe, dessen relative Häufigkeit p nicht übersteigt, dar. Bei den gegebenen Annahmen ist es natürlich, diese Treppenfunktionen mit glatten stetigen Funktionen v(p) und F(p) zu approximieren, ähnlich wie man eine diskrete empirische Verteilungsfunktion üblicherweise durch eine glatte Approximation ersetzt. Aus der Definition folgt unmittelbar:

$$\sum_{p'<p_i^N\le p''} \frac{F^N(p_i^N)-F^N(p_{i-1}^N)}{p_i^N} = \int_{p'}^{p''} \frac{dF^N(p)}{p} = v^N(p'')-v^N(p'), \qquad (5)$$

wonach die Beziehung zwischen p_i^N und $F^N(p)$ durch die Minimum-Lösung der Ungleichung

$$\int_{p_i^N}^{p_1^N} \frac{dF^N(p)}{p} \le i-1, \qquad i = 1,2,\ldots v, \qquad (6)$$

bestimmt wird.

Wenn F(p) monoton wachsend ist, so kann man als Approximation für p_i^N die Lösungen der Gleichungen

$$\int_{p_i}^{p_1} \frac{dF(p)}{p} = i-1, \qquad p_i \approx p_i^N \qquad (7)$$

verwenden. Diese Funktion kann man auch für die Schätzung der Entropie verwenden:

$$H = -\sum_i p_i^N \ln p_i^N = -\sum_m \frac{m}{N} v_m(N) \ln \frac{m}{N} =$$

$$= -\int_{p_v^N}^{p_1^N} \ln(p)\,dF^N(p) \approx -\int_{p_v}^{p_1} \ln(p)\,dF(p). \qquad (8)$$

Aus der Ungleichung

$$\frac{1}{p''}[F^N(p'')-F^N(p')] \le v^N(p'')-v^N(p') \le \frac{1}{p'}[F^N(p'')-F^N(p')], \qquad (9)$$

die bei $p' < p''$ offensichtlich gilt, finden wir den Zusammenhang zwischen v(p) und F(p), wenn $p' \sim p''$, als

$$\frac{dF(p)}{p} = dv(p); \qquad (10)$$

daraus erhalten wir die Beziehung zwischen der Funktion F(p) und der Anzahl unterschiedlicher Elemente, deren Häufigkeiten in das Intervall $[p',p''] \subset [p_n,p_1]$ fallen:

$$\int_{p'}^{p''} dv(p) = \int_{p'}^{p''} \frac{dF(p)}{p} = v(p'') - v(p'). \qquad (11)$$

Unter der Voraussetzung, daß F(p) in $[p_v,p_1]$ differenzierbar ist, führen wir eine Funktion f(p) so ein, daß

$$f(p)\,dp = dF(p). \qquad (12)$$

Dann ist

$$\int_{p'}^{p''} f(p)\,dp = \int_{p'}^{p''} dF(p) = F(p'') - F(p') \approx F^N(p'') - F^N(p') =$$

$$= \sum_{p'<\frac{m}{N}\le p''} v_m(N)\frac{m}{N} = P(p'<p\le p''), \qquad (13)$$

d.h. das Integral auf der linken Seite ist gleich der Wahrscheinlichkeit der Wahl eines Elements aus unserer Stichprobe, dessen Häufigkeit zwischen p' und p'' liegt. Diese Gleichung kann man als eine direkte Definition von f(p) ansehen; offensichtlich wird sie relativ umso genauer , je größer das Intervall [p',p''] ist. Die Funktion f(p) bezeichnen wir - in Analogie zu der üblichen Dichte - als die Funktion der strukturellen Dichte der Menge der Häufigkeiten $\{p_i^N\}$.

Da $f(p) \neq 0$ nur im Intervall $[p_v,p_1]$ gilt und außerhalb dieses Intervalls f(p) identisch Null ist, kann man aufgrund von (13) die übliche Bedingung der Normierung als

$$\int_{p_v}^{p_1} f(p)dp = 1 \qquad (14)$$

schreiben.

Unter Verwendung der Funktion der strukturellen Dichte kann man (8) zu

$$H = -\int_{p_v}^{p_1} f(p)\ln(p)dp, \qquad (15)$$

und (10) zu

$$dv(p) = \frac{f(p)}{p}\,dp \qquad (16)$$

umformen. Daraus erhält man leicht die Zahl unterschiedlicher Elemente $_v[p',p'']$, deren Häufigkeiten zwischen p' und p'' liegen:

$$v[p',p''] - 1 = v^N(p'') - v^N(p') = \int_{p'}^{p''} \frac{f(p)}{p}\,dp. \qquad (17)$$

Insbesondere für die Anzahl unterschiedlicher Elemente in der Stichprobe (d.h., für das "Vokabular") erhalten wir

$$v = v[p_n,p_1] \approx \int_{p_v}^{p_1} \frac{f(p)}{p}\,dp + 1. \qquad (18)$$

Mit Hilfe der Funktion der strukturellen Dichte kann man auch (7) als

$$\int_{p_i}^{p_1} \frac{f(p)}{p}\,dp = i - 1; \quad i = 1,2,\ldots,v \qquad (19)$$

schreiben.

Kennt man also eine der drei Funktionen f(p), F(p) oder v(p), so kennt man auch die Häufigkeitsstruktur der Stichprobe. Obwohl man nicht direkt weiß, zu welchem Element welche Häufigkeit gehört (die Analyse der Häufigkeitsstruktur hat nicht die Intention, dieses Problem zu lösen), kann man trotzdem bei zufälliger Überprüfung eines beliebigen Elements, dessen Häufigkeit in dem gegebenen Intervall liegt, sowohl die Wahrscheinlichkeit der Stichprobe [Formel (13)] als auch die Anzahl unterschiedlicher Elemente, deren Häufigkeiten in dem gegebenen Intervall liegen [Formel (14)], abschätzen. Ebenso kann man die Entropie [Formel (8) und (15)] schätzen. Die Formeln (7) und (9) geben die Möglichkeit, zu der traditionellen Vorstellung der Häufigkeitsstruktur als einer nach abnehmenden Häufigkeiten geordneten Menge überzugehen.

Diese Darstellungsart der Häufigkeitsstruktur kann man offensichtlich immer dann anwenden, wenn die Zahl unterschiedlicher Elemente groß ist und die Bedingung (2) erfüllt ist. Im allgemeinen ist sie dann zweckmäßig, wenn die Funktionen f(p), F(p) oder v(p) als elementare Funktionen leicht integrierbar sind.

2

Wir zeigen, daß eine Funktion der strukturellen Dichte des Typs

$$f(p) = \frac{A}{p^{\alpha}} \qquad (A,\alpha \text{ sind Konstanten}) \qquad (20)$$

dieselbe Häufigkeitsstruktur wie die Mandelbrotsche Formel (1) beschreibt.

Setzt man (20) in (19) ein und löst bezüglich p_i, so erhält man

$$p_i = \frac{(A/\alpha)^{1/\alpha}}{(A/\alpha p_1^{\alpha} - 1+i)^{1/\alpha}} . \qquad (21)$$

Vergleicht man (21) mit (1), so sieht man, daß beide Formeln identisch sind, wenn man

$$K = \left(\frac{A}{\alpha}\right)^{1/\alpha} ; \; B = \frac{A}{\alpha p_1^{\alpha}} - 1; \; \gamma = \frac{1}{\alpha} \qquad (22)$$

setzt.

Die Konstante A wird aufgrund der Normierung (14) bestimmt:

$$A_{\alpha} = \frac{1-\alpha}{p_1^{1-\alpha} - p_v^{1-\alpha}} \qquad \text{für } \alpha \neq 1 \qquad (23a)$$

$$A_1 = \frac{1}{\ln(p_1/p_v)} \qquad \text{für } \alpha = 1 \qquad (23b)$$

Damit haben wir nicht nur gezeigt, daß (1) und (20) im Grunde dieselbe Häufigkeitsstruktur beschreiben, sondern wir haben auch die Möglichkeit erhalten, die Konstanten K und B in der Mandelbrotschen Formel als Funktionen der unmittelbar beobachteten Parameter p_1 und p_v der Stichprobe zu schätzen. Die Anzahl unterschiedlicher Elemente in der Stichprobe, berechnet nach (18), erhält man auch als Funktionen eben dieser Parameter:

$$v = \frac{A}{\alpha}\left(p_v^{-\alpha} - p_1^{-\alpha}\right) + 1 . \qquad (24)$$

Der einzige unabhängige Parameter ist also die Größe $\alpha = 1/\gamma$.

Es ist zu bemerken, daß Formel (21), die man aus (20) erhält, nur ein Spezialfall der breiteren Klasse von Häufigkeitsstrukturen ist, die durch (20) beschrieben werden. Obwohl bei $\alpha = 0$ die erhaltenen Beziehungen ihren Sinn verlieren, ist das speziell bei (20) nicht der Fall, denn setzt man sie in (19) ein und löst nach p_i, so findet man die Approximation für die Häufigkeitsfolge als

$$p_i = p_1 e^{-(i-1)(p_1-p_v)} = p_1 e^{p_1-p_v} e^{-i(p_1-p_v)} , \qquad (25)$$

In der quantitativen Linguistik waren solche Häufigkeitsfolgen bisher nicht bekannt[1]. Es hat sich aber gezeigt, daß nach der Formel (25) die Häufigkeiten elementarer Einheiten wie Buchstaben in geschriebenen Texten oder melodische Intervalle in der Musik abnehmen. Als Beispiel hierfür werden in Abb. 1 die Häufigkeiten russischer Buchstaben und in Abb. 2 die Häufigkeiten melodischer Intervalle in Chopins Nocturnen gezeigt. Beide Graphen sind im halblogarithmischen Maßstab dargestellt, wodurch eine exponentielle Funktion in eine Gerade transformiert wird.

Analoge Graphen erhält man für das Englische (mehrere Stichproben), Französische, Althebräische, Rumänische, Estnische, Ungarische und Deutsche. In allen Fällen zeigte sich bei den seltensten Buchstaben eine starke Abweichung nach unten von der Kurve (25). Eine allgemeine Häufigkeitsgrenze, unterhalb welcher Abweichungen vorkamen, kann als 0.01 ± 0.005 angesetzt werden. Einige gemeinsame Abweichungen von (25) zeigten sich beim Georgischen, Armenischen, Suaheli und Hausa. Ein charakteristisches Beispiel für die Häufigkeit georgischer Buchstaben ist auf Abb. 3 dargestellt. Auch in diesen Fällen nimmt die Häufigkeitsmasse

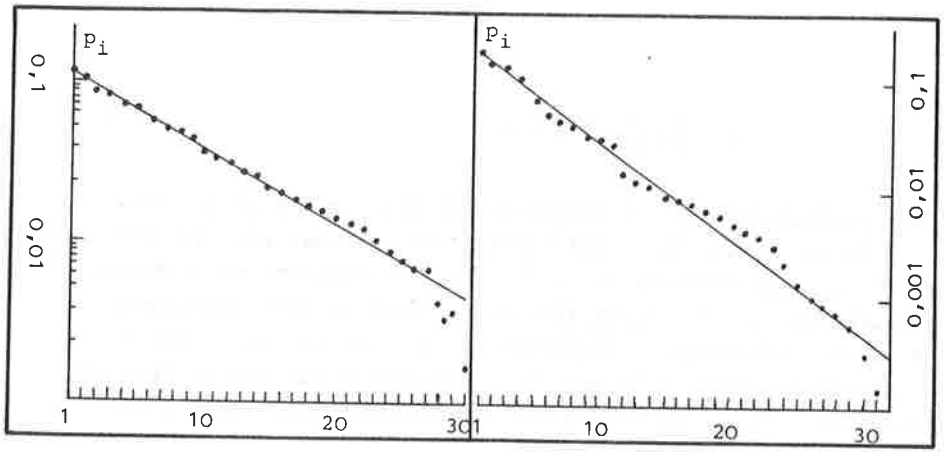


Abb. 1. Buchstabenhäufigkeiten in russischen Texten (nach Lebedev und Garmaš; Zwischenraum ausgelassen). Die unterbrochene theoretische Gerade wurde laut (25) berechnet, wobei p_1 und p_V aus den Daten eingesetzt wurden

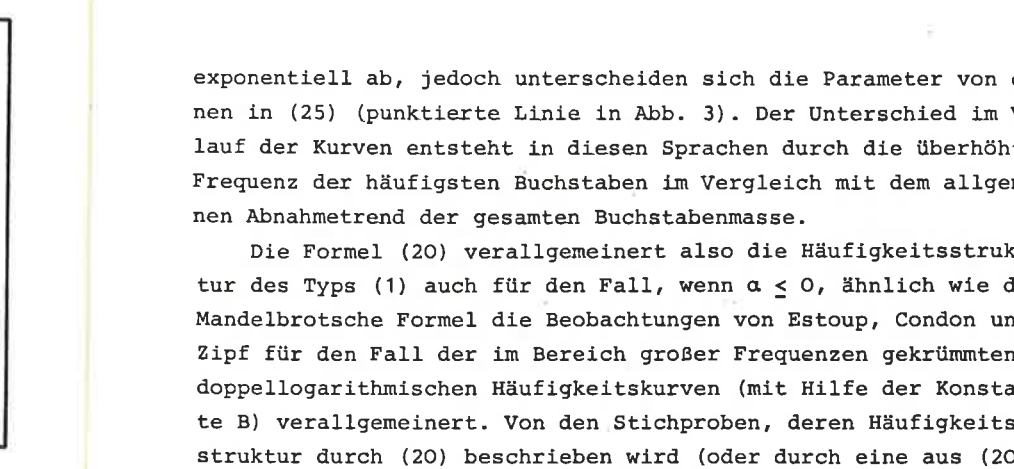


Abb. 2. Häufigkeiten melodischer Intervalle in Nokturnen von F. Chopin (eigene Zählungen zusammen mit M.G. Boroda)

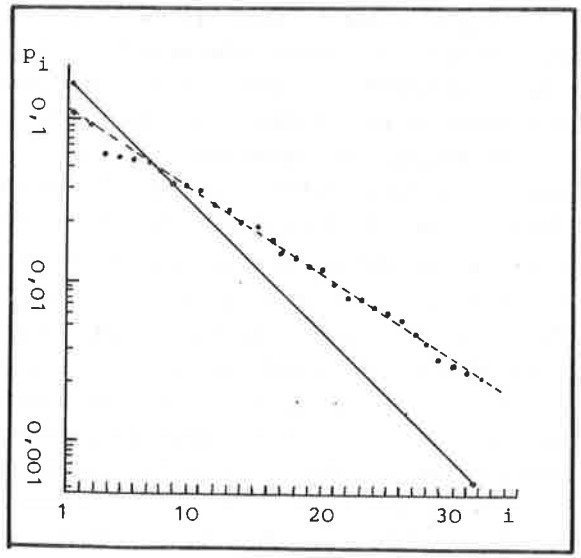


Abb. 3. Buchstabenhäufigkeiten in georgischen Texten (nach Gačečiladze & Eliašvili 1958)

exponentiell ab, jedoch unterscheiden sich die Parameter von denen in (25) (punktierte Linie in Abb. 3). Der Unterschied im Verlauf der Kurven entsteht in diesen Sprachen durch die überhöhte Frequenz der häufigsten Buchstaben im Vergleich mit dem allgemeinen Abnahmetrend der gesamten Buchstabenmasse.

Die Formel (20) verallgemeinert also die Häufigkeitsstruktur des Typs (1) auch für den Fall, wenn $\alpha \leq 0$, ähnlich wie die Mandelbrotsche Formel die Beobachtungen von Estoup, Condon und Zipf für den Fall der im Bereich großer Frequenzen gekrümmten doppellogarithmischen Häufigkeitskurven (mit Hilfe der Konstante B) verallgemeinert. Von den Stichproben, deren Häufigkeitsstruktur durch (20) beschrieben wird (oder durch eine aus (20) abgeleitete Reihe) werden wir sagen, daß sie dem verallgemeinerten Zipf-Mandelbrotschen Gesetz, unabhängig von dem Parameter α, folgen.

3

Von den allgemeinen Beziehungen gehen wir jetzt zu der Untersuchung eines spezifischen Falles über, der z.B. für lexikalische Stichproben charakteristisch ist. Eine wichtige Eigenschaft dieser Stichproben ist ein obligatorisch großer Anteil von hapax legomena und anderer umfangreicher Klassen von seltenen Wörtern. Dieser Umstand wurde zum ersten Mal von Yule (1944) bemerkt; auch später wurden keine Stichproben ohne hapax legomena beobachtet. Mit anderen Worten, die Stichproben schöpfen ihr "potentielles Vokabular" nicht aus, und die Wachstumskurve des Vokabulars erreicht den Sättigungsbereich nicht.

Das obligatorische Vorkommen einer großen Klasse von hapax legomena erlaubt zu schreiben

$$p_v = \frac{1}{N}. \tag{26}$$

Aufgrund dieser Gleichheit kann man die Konstanten A,K,B sowie den Wortschatz auf den Umfang derjenigen Stichprobe, deren Häufigkeitsstruktur dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgt, in Bezug setzen. Im weiteren werden wir den Umfang einer solchen Stichprobe, die laut Definition diesem Gesetz folgt, mit Z bezeichnen und den "Zipfschen Umfang" nennen.

Setzt man (26) in (23) ein, so erhält man

$$A_\alpha = \frac{1-\alpha}{p_1^{1-\alpha} - Z^{\alpha-1}}; \quad A_1 = \frac{1}{\ln(Zp_1)}. \tag{27}$$

Weitere Einsetzungen ergeben

$$K_\alpha = \left[\frac{1-\alpha}{\alpha(p_1^{1-\alpha} - Z^{\alpha-1})}\right]^{1/\alpha}; \quad K_1 = \frac{1}{\ln(Zp_1)}; \tag{28}$$

$$B_\alpha = \frac{1}{\alpha p_1^\alpha (p_1^{1-\alpha} - Z^{\alpha-1})} - 1; \quad B_1 = \frac{1}{p_1 \ln(Zp_1)} - 1. \tag{29}$$

Der Ausdruck (21) wird in dem Fall, daß $\alpha = 1$ ist, zu

$$p_i = \frac{\frac{1}{\ln(Zp_1)}}{\frac{1}{p_1 \ln(Zp_1)} - 1 + i}. \tag{30}$$

Aus (24) erhalten wir die Zahl unterschiedlicher Elemente (das "Vokabular") als die Funktion des Zipfschen Umfangs:

$$v^{(\alpha)}(Z) = \frac{A}{\alpha}(Z^\alpha - p_1^{-\alpha}) + 1. \tag{31a}$$

Im speziellen Fall, wenn $\alpha = 1$ ist, und die Größe

$$\frac{1}{p_1 \ln(Zp_1)} - 1 = B$$

vernachlässigt wird, kann man schreiben

$$v^{(1)}(Z) \approx \frac{Z}{\ln(Zp_i)} \tag{31b}$$

Da sich im untersuchten Fall selten vorkommende Elemente in umfangreiche Klassen mit der Häufigkeit $m = 1,2,\ldots$ gruppieren, bestimmen wir aus (17) die Anzahl unterschiedlicher Elemente, die jeweils die absolute Häufigkeit m haben, als die Anzahl der Elemente, deren Häufigkeit im halboffenen Intervall der relativen Häufigkeiten $[\frac{m}{Z}, \frac{m+1}{Z})$ liegt, als

$$v_m(Z) = v[\tfrac{m}{Z}, \tfrac{m+1}{Z}) = A \int_{\frac{m}{Z}}^{\frac{m+1}{Z}} \frac{dn}{p^{\alpha+1}}. \tag{32}$$

Für $\alpha \neq 1$ ergibt sich daraus

$$v_m^{(\alpha)}(Z) = v^{(\alpha)}(Z)\left(\frac{1}{m^\alpha} - \frac{1}{(m+1)^\alpha}\right). \tag{33a}$$

Für $\alpha = 1$ ergibt sich

$$v_m(Z) = \frac{v(Z)}{m(m+1)}. \tag{33b}$$

Obwohl die Formeln (33a,b) in der Literatur noch nicht vorkommen, müssen wir bemerken, daß die allgemeine Abhängigkeit der Abnahme des Klassenumfangs von der Häufigkeit der in ihr enthaltenen Elemente dieselbe ist wie in den diesbezüglichen Ausdrücken in den Arbeiten von Yule (1944), Kalinin (1965) und Booth (1967). Formel (33b) ist ein Spezialfall der Formel von Simon (1960). Auf diese Weise widersprechen die erhaltenen Beziehungen nicht den bekannten Verallgemeinerungen.

4

Bis jetzt war unsere Darstellung rein formal. Es wurden keine Voraussetzungen über statistische oder kombinatorische Mechanismen verwendet, durch die die analysierten Strukturen erzeugt werden. Obwohl wir auch im weiteren auf Hypothesen über die Struktur der Verteilung der "wahren Wahrscheinlichkeiten" verzichten werden (die Gründe hierfür werden am Schluß diskutiert), führen wir ein einfaches kombinatorisches Modell ein, das uns ermöglicht, die Dynamik der Veränderung der Häufigkeitsstruktur und die Zunahme der Anzahl unterschiedlicher Elemente in Stichproben des lexikalischen Typs, für die (26) gilt, zu beschreiben.

Es soll ein unendlich großer statistisch homogener Text gegeben sein. Unter einem homogenen Text verstehen wir laut Kalinin (1965) einen solchen Text, in dem die Wahrscheinlichkeit der Verwendung jedes Wortes von seiner Position und von den früher verwendeten Wörtern unabhängig ist (d.h. wir setzen nur die Existenz der Wahrscheinlichkeiten voraus, ohne jegliche Beschränkungen über ihre Art; dies hängt damit zusammen, daß unterschiedliche Verteilungen zu identischen mathematischen Erwartungen der selten vorkommenden Wörter führen). Genaugenommen wird diese Voraussetzung in realen Texten nicht erfüllt, aber sie erlaubt, Zusammenhänge, die mit der Realität gut übereinstimmen, abzuleiten. Unser Modelltext soll noch zwei weitere Bedingungen erfüllen:

1. In einer Stichprobe beliebiger Länge aus diesem Text soll es hapax legomena geben, d.h. im ganzen Verlauf des Textes gilt die Beziehung (26); mit anderen Worten, das Vokabular des Textes ist unendlich.

2. In Stichproben mit festem Umfang Z aus diesem Text soll das verallgemeinerte Zipf-Mandelbrotsche Gesetz gelten; d.h. die Wahrscheinlichkeit häufiger Wörter folgt der Beziehung (1) mit Koeffizienten, die durch (28)-(29) gegeben sind, und die Ausdrücke (33) betrachten wir als mathematische Erwartungen der Zahl der m-maligen Wörter in solchen Stichproben. So wird anstatt der Vorgabe einer nichtbeobachteten Verteilung im Bereich

selten vorkommender Wörter die beobachtete Struktur in einem fixierten Stichprobenumfang postuliert.

Untersuchen wir, wie die Häufigkeitsstruktur in Stichproben mit einem anderen, von Z unterschiedlichen Umfang N gestaltet wird. Zwecks Anschaulichkeit führen wir zuerst eine qualitative Analyse durch.

Wenn sich der Umfang einer Stichprobe ändert, so bleiben die Häufigkeiten häufiger Wörter unverändert (bis auf eine rein statistische Streuung, die man mit üblichen Methoden, z.B. mit dem t-Test beurteilen kann). Jedoch wird das letzte Element in der empirischen Menge der Häufigkeiten diesmal nicht 1/Z sondern 1/N. Das heißt, wenn $N < Z$, dann hat diese Menge weniger Elemente als die Menge mit Zipfschem Umfang, und wenn $N > Z$, dann ist es umgekehrt. Offensichtlich gilt also: wenn alle Häufigkeiten nach wie vor auf der Kurve (1) liegen, dann wird die Bedingung der Normiertheit verletzt.

Nehmen wir an, daß wir die ursprüngliche Stichprobe verringert haben, so daß der neue Umfang $N < Z$ ist. Da sich die Häufigkeiten häufiger Wörter nicht geändert haben, so müssen, zwecks Normierung, wegen des höheren Wertes des letzten Elements der Häufigkeitsfolge (und offensichtlich auch wegen der kleineren Zahl der Elemente), die Häufigkeiten seltener Wörter größer sein als derjenigen mit derselben Rangzahl im Zipfschen Umfang. Vergrößert man den Umfang der Stichprobe, so erhält man auf Grund derselben Überlegung das umgekehrte Bild: Der "Schweif" der Kurve im Bereich seltener Wörter muß unterhalb der ursprünglichen Zipfschen Kurve liegen, d.h. die Häufigkeitsstruktur unseres Modelltextes kann nicht stabil sein; sie scheint eine Funktion des Stichprobenumfangs zu sein.

Diese qualitative Überlegung kann durch eine kombinatorische Berechnung unterstützt werden.

Kalinin (1965) hat das Problem der Deformierung der Häufigkeitsstruktur in homogenen Stichproben gelöst. Wenn die mathematischen Erwartungen des Vokabulars $v(N_o)$ und der Zahl der m-maligen Wörter $v_m(N_o)$ in einer "Basisstichprobe" mit Umfang N_o aus demselben Text bekannt sind, dann sind in Stichproben

eines beliebigen Umfangs N das Vokabular $v(N,N_o)$ und die Zahl der m-maligen Wörter $v_m(N,N_o)$ identisch:

$$v(N,N_o) = v(N_o) - \sum_{j=1}^{\infty} \left(1 - \frac{N}{N_o}\right)^j v_j(N_o); \tag{34}$$

$$v_m(N,N_o) = (-1)^{m+1} \frac{N^m}{m!} \frac{d^m}{dN^m} v(N,N_o) \tag{35}$$

Setzt man die Häufigkeitsstruktur mit Zipfschem Umfang (31) und (33) als Charakteristikum der Basisstichproben in (34) und (35) ein, so erhält man

$$v(N,Z) = v(Z)\frac{1-x}{x} \lambda(x,\alpha); \tag{36}$$

$$v_m(N,Z) = v(Z)\frac{(-1)^{m+1}}{m!} \left(\frac{N}{Z}\right)^m \frac{d^m}{dx^m} \left(\frac{1-x}{x} \lambda(x,\alpha)\right) \tag{37}$$

wo $x = 1 - \frac{N}{Z}$

und $\lambda(x,\alpha) = \sum_{j=1}^{\infty} \frac{x^j}{j^\alpha}$.

Die Reihe $\lambda(x,\alpha)$ hat den Konvergenzradius $|1|$; ihre analytische Fortsetzung im Intervall $-\infty<x<1$ ist

$$\lambda(x,\alpha) = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \frac{t^{\alpha-1} x e^{-t}}{1-xe^{-t}} dt, \tag{38}$$

wo $\Gamma(\alpha)$ die Gamma-Funktion mit dem Argument α ist.

Wenn $\alpha = 1$, so kann man $v(N,Z)$ und $v_m(N,Z)$ mit Hilfe elementarer Funktionen ausdrücken:

$$v(N,Z) = v(Z) \frac{\ln \frac{N}{Z}}{\frac{Z}{N} - 1}; \tag{39}$$

$$v_m(N,Z) = v(Z) \frac{(-1)^{m+1}}{m!} \left(\frac{N}{Z}\right)^m \frac{d^m}{dx^m} \left(\frac{x-1}{x} \ln(1-x)\right). \tag{40}$$

Bei kleinem m bekommt man speziell:
für hapax legomena:

$$v_1(N,Z) = v(Z) \frac{M-1}{X-1}; \tag{41}$$

für zweimal vorkommende Wörter:

$$v_2(N,Z) = v(Z) \frac{X+1-2M}{2(X-1)^2}; \tag{42}$$

für dreimal vorkommende Wörter:

$$v_3(N,Z) = v(Z) \frac{6M+X^2-5X-2}{6(X-1)^3}, \tag{43}$$

wo $X = \frac{Z}{N}$

und $M = \frac{\ln X}{1-\frac{1}{X}}$ ist.

Wenn $X \approx 1$, dann kann man den Ausdruck

$$v_m(N,Z) = v(Z) \sum_{i=m}^{\infty} \frac{(1-X)^{i-m}}{i(i+1)} \tag{44}$$

verwenden.

Formel (40) wird bei großem m sehr umfangreich und schwer auswertbar. Daher kann man den Verlauf des Bereichs mit kleinen Häufigkeiten auf folgende Weise leichter beschreiben.

Ist der Wert der absoluten Häufigkeit m gegeben, dann kann man die letzte Rangzahl in der allgemeinen Häufigkeitsliste für die Wörter mit der Häufigkeit m + 1 [die rechte Abszisse des Rechtecks für die Wörter mit der Häufigkeit m + 1] folgendermaßen berechnen

$$i_{m+1} = v(N,Z) - \sum_{j=1}^{m} v_j(N,Z) = v(Z) \sum_{j=1}^{\infty} \frac{(1-X)^{j-1}}{j+m}. \qquad (45a)$$

Diesen Ausdruck kann man für Werte von $x \approx 1$ benutzen. Im allgemeinen Fall soll man den Ausdruck

$$i_{m+1} = \frac{v(Z)}{(1-X)^m} \left[\frac{\ln X}{X-1} - \sum_{k-1}^{m} \frac{(1-X)^{k-1}}{k} \right] \qquad (45b)$$

benutzen, woraus folgt

$$v_m(N,Z) = i_m - i_{m+1}.$$

Die linke Abszisse des Rechtecks für Wörter mit Häufigkeit m ist dann natürlich $i_{m+1}+1$. Den Ausgangspunkt für die Konstruktion der theoretischen Graphen bildet das vollständige Vokabular des Textes $v(N,Z)$, das die von hapax legomena gebildete Abszisse des untersten Rechtecks bestimmt. Wenn man den Graphen in relativen Häufigkeiten zeichnet, dann ist die Ordinate dieses Punktes gleich 1/N.

Auf Abb. 4 sieht man die Veränderung der Häufigkeitsstruktur des Textes bei der Veränderung des Stichprobenumfangs; hierbei wurde $\alpha = 1$ gesetzt.

Beim Umfang Z wird die Häufigkeitsfolge durch (21) beschrieben. Die ununterbrochene, fallende Kurve ist die Häufigkeitsstruktur beim Umfang Z = 27000 und $p_1 = 0.07$. Die "Treppen" im unteren Teil des Graphen sind laut (45a und b) für verschiedene Werte des Umfangs N berechnet worden. Bei wachsendem m nähern sich die Treppen der ursprünglichen Kurve, die für den Umfang Z

laut (30) berechnet wurde. Das vollständige Vokabular bei diesem Umfang ist gleich $v(Z)$. Die kleine Krümmung im oberen Teil des Graphen wird durch den Umstand verursacht, daß $B > 0$. Je mehr sich der Stichprobenumfang N von dem Zipfschen Umfang Z unterscheidet, desto stärker ist der "Schweif" der Kurve gekrümmt.

Beim Umfang $N_1 < Z$ liegt der größte Teil der "Rechtecke" über der ursprünglichen Kurve (21). Wenn man die Größe γ direkt für diesen Graphen bestimmen wollte, so würde sie im unteren Teil des Graphen etwas kleiner als 1 ausfallen. Die größere Länge der niedrigsten Rechtecke (m = 1,2) zeugt von der Vergrößerung des Anteils seltener Wörter im Vokabular (es ist zu bemerken, daß bei $N < Z$ der Anteil der hapax legomena über die Hälfte des Wortschatzes beträgt).

Vergrößert man die Stichprobe (Umfang N_2 ist unwesentlich kleiner als Z), so verlängert sich die Folge von Häufigkeiten (d.h. das Vokabular wächst) und die rechten Ecken der Rechtecke nähern sich der Kurve (21). Wenn N = Z, so wird (21) zur oberen Grenze dieser Rechtecke; damit der Graph nicht unübersichtlich wird, haben wir diese Situation nicht gezeichnet, man kann sie sich aber leicht durch $v(N,Z) \to v(Z)$ bei $N \to Z$ vorstellen. Die Anzahl der hapax legomena nähert sich dabei der Hälfte des Vokabulars.

Bei $N_3 > Z$ beginnt der Schweif unter der Kurve (21) zu liegen. Der Anteil seltener Wörter im Vokabular nimmt ab (obwohl ihre absolute Zahl wächst); die Anzahl der hapax legomena wird kleiner als die Hälfte des Vokabulars. Bei der empirischen Schätzung von γ einer derartigen Kurve ist es nicht leicht zu bestimmen, wo ihr geradliniger Teil endet; die dargestellte Krümmung würde zu dem Schluß führen, daß das Zipfsche Gesetz in dem gegebenen Text schlecht erfüllt ist und daß γ im unteren Teil des Graphen etwas größer als 1 ist.

Schon bei einem flüchtigen Blick auf Abb. 4 sieht man also, daß das vorgeschlagene Modell Situationen wiedergibt, die man in empirischen Graphen üblicherweise vorfindet. Aus dieser Analyse folgt, daß die "Abweichungen im Schweif" eine gesetzmäßige und unvermeidliche Erscheinung sind in dem Fall, wo sich der Stichprobenumfang vom Zipfschen Umfang Z unterscheidet. D.h.,

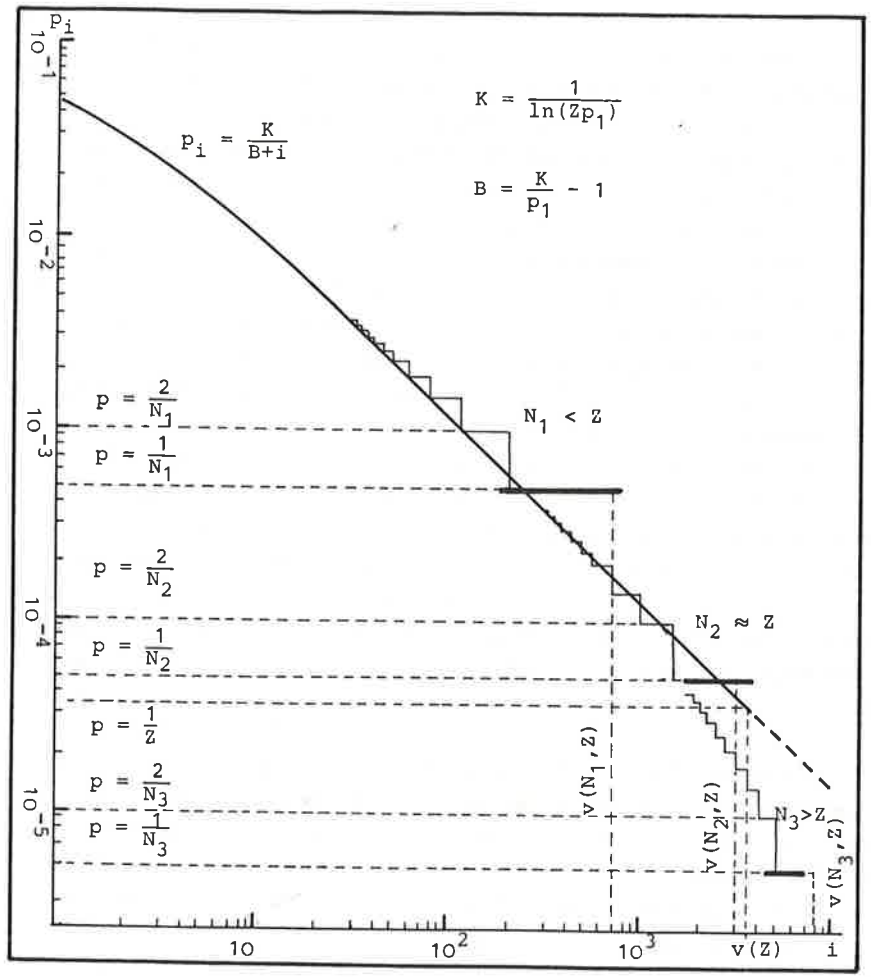


Abb. 4. Das allgemeine Bild der Dynamik der Häufigkeitsstruktur

ein beliebiger homogener Text kann dem Zipf-Mandelbrotschen Gesetz in seiner "kanonischen" Form (1) nur bei einem einzigen Wert des Stichprobenumfangs folgen (wenn er ihm überhaupt folgen kann). Bei allen anderen Umfängen entstehen gesetzmäßige Krümmungen der Treppenfunktion, ungeachtet dessen, daß die gegebenen Werte von α oder γ gleich bleiben. Die Kenntnis des Z-Wertes erlaubt es, sowohl das Wachsen des Vokabulars innerhalb des Textes [Formeln (37) und (39)] als auch seine Häufigkeitsstruktur beim beliebigen Umfang [Formeln (40-45) und (37)] zu beschreiben.

Die durch (36) - (45) beschriebenen Häufigkeitsstrukturen kann man als eine natürliche Verallgemeinerung des Zipf-Mandelbrotschen Gesetzes auf den Fall, wo der Stichprobenumfang N beliebig und ungleich Z ist (d.h. ungleich dem einzigen Umfang, bei dem die Beziehungen (27) - (33) "konstruktionsgemäß" erfüllt sind) betrachten. Diese Häufigkeitsstrukturen kann man als "Quasizipfsche" bezeichnen, da sie die Folge des Umstandes sind, daß beim Umfang Z das Zipf-Mandelbrotsche Gesetz erfüllt wird.

Klären wir die besondere Rolle des Wertes Z, die mit dem allgemeinen Verlauf des Vokabularwachstums innerhalb des Textes zusammenhängt. Betrachten wir zwei Texte mit identischen p_1 und α, aber unterschiedlichen Zipfschen Umfängen $Z_1 > Z_2$. Erhebt man aus beiden Texten gleiche Stichproben mit dem Umfang N, so sieht man, daß

$$v(N,Z_1) > v(N,Z_2). \qquad (46)$$

Das heißt, ein Vokabular wächst beim Anwachsen des Stichprobenumfangs umso schneller, je größer die Größe Z . Das bedeutet, daß die relative Sättigung des Textvokabulars nicht nur von $\alpha = 1/\gamma$, sondern auch vom Wert des Zipfschen Umfangs abhängt. Da man im Rahmen einer Sprache p_1 und α als Konstanten betrachten kann, kann man Z als ein akzeptables Maß des relativen Vokabularreichtums betrachten.

5

Als Ergebnis der Untersuchung erhielten wir statt der Konstanten K,B und γ in der Mandelbrotschen Formel die Parameter p_1, p_v und α oder (im Fall kleiner lexikalischer Stichproben, die hapax legomena enthalten) p_1', Z und α. Die Parameter p_1 und p_v kann man direkt aus den Daten der Stichprobe abschätzen; α = = 1/γ; der Zipfsche Umfang Z scheint eine neue Größe zu sein, die in bisherigen linguostatistischen Untersuchungen nicht vorhanden war. Zipf (1949) hat sie zwar unter dem Namen "optimaler Umfang" in seiner Antwort auf die Kritik von M. Joos (1936) eingeführt, (Joos behauptete, daß die von Zipf untersuchte Gesetzmäßigkeit kein universales Gesetz sein kann), aber sonst wurde diese Idee nicht weiter entwickelt. Als Zipf versuchte, den Wert dieses Umfangs empirisch zu bestimmen, hat man ihn sogar beschuldigt, daß er "das Experiment der Theorie anpaßt" (vgl. Frumkina 1961).

Der Begriff des Zipfschen Umfangs erlaubt es, bei kleinen Stichproben den Mechanismus der "Abweichung im Schweif" von "kanonischen" Kurven des Typs (1) im Bereich der kleinen Häufigkeiten zu verstehen und die Größe dieser Abweichung abzuschätzen. Dies ist das wesentliche theoretische Resultat der vorliegenden Arbeit.

Der Vergleich mit empirischem Material zeigt, daß man lexikalische Stichproben allgemein mit dem vorgelegten Modell bei α = 1 unabhängig von Stil, Genre und Sprache beschreiben kann. Dabei kommt der Umfang Z dem Umfang eines größeren literarischen Werkes sehr nahe (vgl. Orlov 1969a,b,1970), denn sowohl das Vokabular solcher Texte als auch ihre Häufigkeitsstruktur lassen sich mit Hilfe von (30), (31b) und (33b) zufriedenstellend (in der Regel mit einer Fehlergenauigkeit von ± 20%) beschreiben, wenn man für Z den reellen Textumfang einsetzt. In Orlov (1969b) wurden ergänzend die Veränderungen des Vokabulars und der Häufigkeitsstruktur innerhalb des Textes aufgrund der aus ihm erhobenen Stichprobe analysiert. Die Veränderungen entsprechen den Ausdrücken (39) und (41-43) mit derselben Genauigkeit.

In Nadarejšvili & Orlov (1971) wurde der Unterschied der Zuwachsrate des Vokabulars in Texten unterschiedlicher Länge, die von demselben Autor stammen, gezeigt; dort wurde auch gezeigt, daß sehr große Texte in ihren vom Autor bestimmten Teilen (Kapitel, Band usw.) dem Zipf-Mandelbrotschen Gesetz folgen können.

Gleichzeitig gilt aber, daß willkürliche Stichproben (Ausschnitte und Zusammenstellungen vieler Texte in eine Stichprobe, die die "Sprache als ganze" oder eine "Fachsprache" repräsentieren sollen) in der Regel dem kanonischen Zipf-Mandelbrotschen Gesetz (1) nicht folgen. Die Abweichungen des Vokabulars vom Umfang (31b) erreichen 50 - 100%, der Verlauf der Häufigkeitskurve stimmt mit der theoretischen Kurve (30) nicht überein, die Krümmung der Treppenkurve im Schweif des Graphen ist erheblich (vgl. Orlov 1974). Ausnahmen, d.h. Fälle, wo die kanonische Form genau erfüllt wird, sind selten und können durch die zufällige Nähe des Stichprobenumfangs zu dem Zipfschen Umfang erklärt werden. Jedoch erlaubt das entwickelte Modell praktisch in allen Fällen, die Parameter solcher Stichproben miteinander in Zusammenhang zu bringen (vgl. Orlov 1976).

Als Beispiel bringen wir in der Tabelle 1 die Resultate der Berechnung des Anwachsens des Vokabulars und der Zahl der hapax legomena mit Hilfe des dargestellten Modells für die Daten von Guiraud (1960). In der Tabelle werden auch die von Kalinin (1965) berechneten theoretischen Werte angegeben. Kalinin geht bei den Berechnungen von der Größe des Vokabulars in einem beliebigen Punkt N_o und von den in relativ kleinen Abständen genommenen sukzessiven Anzahlen der hapax legomena aus (deswegen konnte Kalinin die theoretische Prognose für die Umfänge N = 100000 und N = 200000 nicht berechnen). Bei unseren Berechnungen benutzten wir den Vokabularumfang v = 3040 bei einem Stichprobenumfang N = 200000. Diese Werte wurden in die Gleichung

$$v(20000, Z) = 3040$$

eingesetzt. Die approximative Lösung dieser Gleichung[2)] bezüg-

Tabelle 1. Vergleich der Prognosen

Daten von Guiraud			Prognose von Kalinin			Prognose laut (39) und (42); Z=27000, p_1=0,07			
N	v	v_1	Berechnet bei	Theoretisches Vokabular v(N)	$\frac{N}{v(N)}$	v(N,Z)	$\frac{v}{v(N,Z)}$	$v_1(N,Z)$	$\frac{v_1}{v_1(N,Z)}$
2000	800	600	N_o = 2000	750	1,07	746	1,07	518	1,16
4000	1250	800	N_o = 4000	1300	0,96	1190	1,05	815	0,98
6000	1580	950	- " -	1660	0,95	1540	1,03	950	1,00
8000	1900	1070	- " -	1950	0,97	1830	1,04	1100	0,98
10000	2120	1160	- " -	2200	0,96	2090	1,01	1215	0,95
12000	2320	1240	- " -	2420	0,96	2320	1,00	1320	0,94
14000	2505	1320	- " -	2610	0,96	2530	0,99	1407	0,94
16000	2710	1390	- " -	2800	0,97	2720	1,00	1490	0,93
18000	2775	1450	- " -	2960	0,94	2900	0,96	1547	0,96
20000	3040	1500	N_o = 10000	3060	0,99	3070	-	1637	0,92
30000	3650	1710	- " -	3730	0,98	3740	0,98	1790	0,96
40000	4200	1880	- " -	4270	0,98	4320	0,97	2000	0,94
50000	5000	2100	- " -	4700	1,06	4800	1,04	2145	0,98
100000	6000	2350	-	-	-	6240	0,96	2475	0,95
200000	7500	2750	-	-	-	8270	0,91	2850	0,96

lich Z ergab Z ≈ 27000 (d.h. wir fanden diesen Zipfschen Umfang für den Modelltext, der bei einem Stichprobenumfang von N = 20000 den tatsächlich beobachteten Vokabularumfang ergibt; mit anderen Worten: betrachtet man (39) als die Gleichung für die Kurve des Vokabularwachstums, die von dem Parameter Z abhängt, dann setzt man diese Kurve mit einem experimentell beobachteten Punkt gleich). Als p_1 wurde 0.07 angenommen, was für Sprachen mit Artikeln typisch ist (bei einer schwachen Abhängigkeit der benutzten Beziehungen von p_1 ist es möglich, auch "aus der Luft gegriffene" Werte zu benutzen). Aus diesen zwei Zahlen konnte man sowohl den Vokabularumfang als auch die Zahl der hapax legomena in allen Punkten berechnen. Abweichungen von über 5% im Vokabularumfang wurden nur an den äußersten Enden des untersuchten Materials bei zehnfachen Prädiktionen und Retrodiktionen von dem Ausgangspunkt beobachtet (vgl. Spalten 8 und 10, Zeilen für N = 2000 und N = 20000 in Tab. 1).

Der Wert Z, der aus der Analyse dieses Materials folgte, diente als Grundlage für die Berechnung der Häufigkeitsstruktur, wie sie in Abb. 4 dargestellt wird. Sie stellt dadurch die "Restaurierung" der Häufigkeitsstruktur des Materials von Guiraud bei verschiedenen Stichprobenumfängen dar: N_1 = 2000, N_2 = 20000, N_3 = 200000. Starke horizontale Linien am Ende jeder "Treppe", die leicht von der theoretischen Grenze abweichen, stellen empirisch beobachtete "Rechtecke" der hapax legomena dar. Wie aus der Tabelle ersichtlich ist, sollte bei allen dazwischenliegenden Umfängen die Übereinstimmung der theoretischen und experimentellen Daten besser sein als bei diesen extremen Umfängen.

Wie man außerdem in der Tabelle erkennt, liegen die Prognosen für hapax legomena im Durchschnitt etwas höher (ungefähr um 4%); berücksichtigt man diese Verschiebung, dann ist die Unstimmigkeit auch unbedeutend. Kalinins Prognosen erweisen sich trotz des großen Umfangs der verwendeten Ausgangsinformation als weniger genau und als weniger "weitreichend".

Die Hypothese also, daß bei den Stichproben von 27000 Wörtern aus dem Material von Guiraud das Zipf-Mandelbrotsche Gesetz in der "kanonischen" Form (30) gilt (d.h. daß für dieses Mate-

rial der Zipfsche Umfang = 27000 ist), erlaubt es, die Kurve des Vokabularwachstums und die Anzahl der hapax legomena zufriedenstellend zu berechnen. In der vorliegenden Arbeit streben wir nicht nach einer detaillierten Überprüfung des vorgeschlagenen Modells am empirischen Material (das dargelegte Beispiel dient nur zur Illustrierung des Funktionierens des Modells). Weitere Entwicklungen zum Zweck der Beschreibung der Häufigkeitsstruktur des Lexikons und zahlreiche Vergleiche mit empirischen Daten kann man in Orlov (1976) finden. Wie aus den Resultaten dieser Arbeit ersichtlich ist, erlaubt der Wert $\alpha = 1$ die Häufigkeitsstruktur und das Vokabularwachstum in sehr heterogenem lexikalischen Material, dessen Häufigkeitsgraphen manchmal sehr gekrümmt sind, zufriedenstellend zu beschreiben.

Man kann also annehmen, daß α eine Konstante ist, die mit dem Typ der linguistischen Einheiten zusammenhängt (0 für Buchstaben, 1 für Wörter), während Z für einen konkreten Text oder eine Textauswahl charakteristisch ist. Etwas verallgemeinernd kann man sagen, daß α eine Charakteristik einer sprachlichen Ebene (oder, im weiteren Sinne, eines Informationssystems, eines Kodes, der langue), und Z eine Charakteristik der parole (der Nachricht) ist. Der früher nicht entdeckte Zusammenhang dieser Charakteristika mit dem Stichprobenumfang, der zur "Abweichung des Schweifs" führte, diskreditierte die beobachteten Gesetzmäßigkeiten und erlaubte es nicht, von ihnen ausgehend begründete quantitative Schätzungen durchzuführen.

Die Frage, welchen Wert α für andere linguistische und informationstragende Einheiten wie Phonem, Silbe, Digram, Trigram, zwei- und dreigliedrige Wortgruppen usw. annimmt, bleibt vorläufig offen.[3)] Es sind umfangreiche Untersuchungen nötig, aber es ist offensichtlich, daß es nur unter Berücksichtigung der in dieser Arbeit beschriebenen Dynamik der Häufigkeitsstruktur möglich sein wird, sich in der Vielfalt der empirischen Häufigkeitskurven zurechtzufinden. Vorläufige Rechnungen zeigten auch die prinzipielle Adäquatheit des gegebenen Modells für die Beschreibung der Anteile der Farbflächen in malerischen Kunstwerken (vgl. Vološin&Orlov 1972); eine Erscheinung, die der Erfüllung des "kanonischen" Zipf-Mandelbrotschen Gesetzes in abgeschlossenen literarischen Texten vollkommen analog ist, wurde auch in musikalischen Texten beobachtet (vgl. Boroda 1974). Alles dies bezeugt sowohl die außerordentliche Wichtigkeit als auch die außerordentliche Verbreitung des verallgemeinerten Zipf-Mandelbrotschen Gesetzes in dem Bereich der Kommunikationsmittel.

6

In der vorliegenden Arbeit haben wir das Problem der Herkunft und der Begründung des kanonischen Zipf-Mandelbrotschen Gesetzes absichtlich nicht berührt, sondern es als gegeben angenommen. Im Rahmen des dargelegten Modells scheint diese Form lediglich ein Spezialfall zu sein, der durch eine breitere Klasse von Häufigkeitsstrukturen mit gekrümmtem "Schweif" (§4) wesentlich verallgemeinert wird. Die eigenartige Tatsache, daß Häufigkeitsstrukturen größerer literarischer und musikalischer Werke ausgerechnet zu dieser kanonischen Form tendieren (vgl. Orlov 1974, Boroda 1974), zwingt uns, das Problem von einer anderen Seite anzuschauen. Es ist nicht möglich, diese Erscheinung durch irgendwelche statistischen Metatheorien zu erklären. Es muß eingesehen werden, daß die Realisierung der kanonischen Form der Häufigkeitsstruktur gerade in der vollen Länge der Werke sehr wichtig für ihre Perzeption ist und daß der Mensch im Laufe der Schöpfung der Texte imstande ist, ihre Häufigkeitsstruktur so zu organisieren, daß die volle Länge des Textes dem Zipfschen Umfang nahe kommt. Eine detaillierte Besprechung dieses Problems findet man in Arapov & Efimova & Šrejder (1975) und Orlov (1974, 1976).

Wenn man sich schon darüber wundert, daß die Häufigkeitsstruktur eines Textes mit seiner Länge korrespondiert, was ja zielgerichtete Bemühungen des Autors voraussetzt, so muß man sich erst recht darüber wundern, daß der Zipfsche Umfang auch bei lexikalischen Stichproben existiert, die nicht von einem einheitlichen organisierenden Willen erzeugt worden sind (wie z.B. die

durch Kolguškin ausgezählten militärischen Texte oder Kalininas russische Texte über Elektronik, [vgl. Orlov (1976)]. Daraus muß gefolgert werden, daß sich die Worthäufigkeiten "von alleine" zu einer der zahlreichen Formen des Zipfschen Gesetzes formieren - mit anderen Worten, es muß einen latenten, rein statistischen Mechanismus geben, der anscheinend sehr eigenartig ist, - und daß die Übereinstimmung des Zipfschen Umfangs mit der Textlänge in künstlerischen Werken durch eine "parametrische Unterordnung" unter die Länge des geplanten Textes, die mit der "Einstimmung" und "Regulierung" dieses Mechanismus in jedem konkreten Fall verbunden ist, zustande kommt. Die äußeren Charakteristika dieses Mechanismus (wie "input-output") werden durch das dargelegte Modell beschrieben, jedoch muß seine innere Konstruktion noch analysiert werden. Dies ist äußerst wichtig, da man darin die rein statistischen Erscheinungen von den störenden "menschlichen" Faktoren trennen muß.

Zum Schluß fassen wir die Resultate der vorliegenden Arbeit kurz zusammen:

1. Es wird eine neue Art der Beschreibung der Häufigkeitsstruktur vorgeschlagen, die sich von der üblichen geordneten Folge von Häufigkeiten unterscheidet. Der grundlegende Vorzug der vorgeschlagenen Beschreibung, d.h. der Verwendung der Funktion der strukturellen Dichte f(p) oder der äquivalenten Funktionen F(p) bzw. v(p) liegt in der Ersetzung des oft schwierigen Summierens von Zahlenfolgen durch eine Integration.

2. Es wurde die verallgemeinerte Form des Zipf-Mandelbrotschen Gesetzes als eine Funktion der strukturellen Dichte (20) formuliert. Dabei hat sich gezeigt, daß bei $\alpha > 0$ diese Funktion dieselbe Häufigkeitsstruktur wie die Mandelbrotsche Formel (1) beschreibt. Bei $\alpha \leq 0$ erhält man Häufigkeitsfolgen, die bisher unbekannt waren, aber einige von ihnen finden ihre Entsprechungen im Informationsprozeß (Buchstaben, melodische Intervalle). Die gemeinsame formale Genese unterschiedlicher Strukturen zeugt offensichtlich von der Einheitlichkeit der Informationsprozesse auf verschiedenen Ebenen. Außer dieser verallgemeinernden Rolle hat die Darstellung der Zipf-Mandelbrotschen Häufigkeitsstruktur in Form einer Funktion der strukturellen Dichte den Vorteil, daß nur ein unabhängiger Parameter $\alpha = 1/\gamma$ übrig bleibt, die restlichen, K und B kann man als Funktionen von p_1 und p_v ausdrücken.

3. Die Untersuchung kleiner Stichproben, in denen, wie in den lexikalischen, immer hapax legomena vorhanden sind, ermöglichte es, die grundlegenden beobachteten Parameter dieser Stichproben (Umfang, Häufigkeitsfolge, Vokabular, Zahl der einmaligen, zweimaligen usw. Wörter) durch die Hilfe des unabhängigen Parameters α miteinander zu verbinden.

4. Die Untersuchung eines unendlichen homogenen Textes, der beim Umfang Z dem Zipf-Mandelbrotschen Gesetz in seiner "kanonischen Form" folgt, hat gezeigt, daß Z der einzige Wert ist, bei dem diese Form vorkommen kann. Bei allen anderen Stichprobenumfängen folgen unvermeidlich "Abweichungen der Schweife". Die Kenntnis des Z-Wertes erlaubt es, sowohl die Dynamik der Veränderung der Häufigkeitsstruktur, als auch das Anwachsen des Vokabulars bei wachsendem Stichprobenumfang zu beschreiben. Es wurde auch die besondere Rolle von Z geklärt, nämlich seine Beziehung zu dem relativen Vokabularreichtum des Textes [Ungleichung (46)].

5. Der Vergleich mit empirischen Daten, der sowohl hier als auch in anderen Arbeiten durchgeführt wurde, hat gezeigt, daß die Häufigkeitsstruktur und das Vokabularwachstum innerhalb des Textes mit Hilfe des dargelegten Modells bei $\alpha = 1$ adäquat beschrieben werden kann. Das Modell ermöglichte es, eine Übereinstimmung in der Häufigkeitsstruktur des Vokabulars mit vollem Umfang zwischen literarischen und musikalischen Texten zu entdecken. Die Berücksichtigung des Einflusses des Zipfschen Umfangs bei kleinen Stichproben läßt hoffen, daß es auch bei anderen linguistischen Einheiten außer den Wörtern gelingt, die universelle Bedeutung von α für jede solche Einheit zu finden.

In dieser Arbeit wurde also ein Modell der Häufigkeitsstruktur des Zipfschen Typs aufgestellt, das gesetzmäßige Deformationen eben dieser Struktur beschreibt. Wenn man eine klare Vorstellung davon hat, daß die kanonische Häufigkeitsstruktur des Typs (1) nur ein Spezialfall einer breiteren Klasse verwandter

Häufigkeitsstrukturen ist, und wenn man die Dynamik der Häufigkeitsstruktur berücksichtigt, so kann man ohne den überflüssigen Empirismus auskommen, der sich beim Konstruieren aller möglichen spitzfindigen Formeln zur Approximation an empirische Daten einstellt.

Zum Schluß möchte sich der Verfasser bei M.V. Arapov, A.L. Brudno, E.M. Dumanis, N.M. Poliektov-Nikoladze, A.R. Chvoles, R.Ja. Čitašvili und Ju. A. Šrejder, deren Bemerkungen, Ratschläge und kollegiale Unterstützung beim Entstehen dieser Arbeit sehr behilflich gewesen sind, herzlichst bedanken.

ANHANG

Die eingeführte stetige Approximation der Häufigkeitsstruktur in Form einer Funktion der strukturellen Dichte erlaubt es, nicht nur die vorhandenen Parameter miteinander zu verknüpfen und die Werte der Koeffizienten K und B in der Mandelbrotschen Formel zu finden, sondern sie ermöglicht es auch, sowohl an die anderen bekannten Darstellungen der Häufigkeitsstruktur anzuknüpfen als auch neue Darstellungsarten zu finden. Als Beispiel erläutern wir kurz zwei Darstellungsformen des Zipf-Mandelbrotschen Gesetzes, die einen direkten Vergleich mit empirischen Daten erlauben.

Die erste Form ist in den Fällen geeignet, wenn man einen Vergleich mit empirischem Material unternehmen muß, das in Form von Tabellen oder Graphiken der sogenannten "Textauffüllung mit dem Vokabular" gegeben ist. Wir führen die Funktion der kummulativen Häufigkeit ein:

$$\varphi(i) = \sum_{j=1}^{i} p_j, \qquad i = 1,2,\ldots,v \tag{47}$$

wo die Menge der empirischen Häufigkeiten $\{p_j\}$, nach abnehmender Größe geordnet ist. Die Größe $\varphi(i)$ stellt die Wahrscheinlichkeit der zufälligen Wahl eines Wortes aus der gegebenen Stichprobe dar, dessen Häufigkeit in dieser Stichprobe nicht kleiner als p_i ist. Wenn $f(p)$ die Häufigkeitsstruktur der Stichprobe beschreibt, dann erhält man aufgrund von (13)

$$\varphi(i) = P(p > p_{i+1}) = \int_{p_{i+1}}^{p_1} f(p)dp. \tag{48}$$

Wenn $f(p)$ durch (20) mit $\alpha = 1$ gegeben ist, dann ist

$$\varphi(i) = \left(\ln \frac{p_1}{p_{i+1}}\right)\left(\ln \frac{p_1}{p_v}\right)^{-1}, \tag{49}$$

wobei man p_i aus (30) erhält.

Ein Beispiel derartiger Darstellung des Zipf-Mandelbrotschen Gesetzes und die Form der theoretischen und empirischen Kurven wurde in Orlov (1974) gebracht.

Die zweite Darstellungsform der untersuchten Gesetzmäßigkeiten ist bisher weder in der Praxis noch in der Theorie der quantitativen Linguistik bekannt; wir werden jedoch zeigen, daß sie einige Vorteile hat.

Wir führen den Begriff einer "geometrischen Häufigkeitsklasse" ein (im weiteren wird das Wort "geometrisch" einfachheitshalber ausgelassen). Dann zerlegen wir das Intervall der Häufigkeiten $[p_v,p_1]$ in eine Folge von disjunkten Teilintervallen, die einander in Punkten $\varphi_j = p_v C^{j-1}$ berühren, wo $C > 1$ eine beliebige Konstante ist. Die Längen solcher Teilintervalle bilden eine geometrische Folge mit der Basis C.

Bei der Häufigkeitsstruktur des Lexikons werden wir diejenigen Elemente (Wörter), die in das j-te Teilintervall $[\rho_j,\rho_{j+1})$ fallen, als die j-te Häufigkeitsklasse, und die Anzahl unterschiedlicher Elemente $v[\rho_j,\rho_{j+1})$ in dieser Klasse als den Umfang der j-ten Häufigkeitsklasse bezeichnen. Die Größe, die die Wahrschein-

lichkeit der zufälligen Wahl des Elements der j-ten Häufigkeitsklasse darstellt,

$$P[\rho_j,\rho_{j+1}) = \sum_{\rho_j \le p_i < \rho_{j+1}} p_i, \qquad (50)$$

bezeichnen wir als das relative Gewicht der j-ten Häufigkeitsklasse, und $N[\rho_j,\rho_{j+1}) = NP[\rho_j,\rho_{j+1})$, wobei N der Stichprobenumfang ist, als das absolute Gewicht dieser Gruppe.

Aus (13), (17) und (20) erhält man leicht

$$v[\rho_j,\rho_{j+1}) = \frac{A}{\alpha}\left(\rho_{j+1}^{-\alpha} - \rho_j^{-\alpha}\right) \qquad (51)$$

$$P[\rho_j,\rho_{j+1}) = \frac{A}{1-\alpha}\left(\rho_{j+1}^{1-\alpha} - \rho_j^{1-\alpha}\right). \qquad (52)$$

Das Verhältnis des Umfangs einer Häufigkeitsklasse zu dem Umfang der vorhergehenden Häufigkeitsklasse findet man als

$$\frac{v[\rho_{j+1},\rho_{j+2})}{v[\rho_j,\rho_{j+1})} = \frac{C^{-\alpha j-2\alpha} - C^{-\alpha j-\alpha}}{C^{-\alpha j-\alpha} - C^{-\alpha j}} = C^{-\alpha}. \qquad (53)$$

Analog folgt für die Gewichte zweier aufeinander folgenden Klassen

$$\frac{N[\rho_{j+1},\rho_{j+2})}{N[\rho_j,\rho_{j+1})} = \frac{P[\rho_{j+1},\rho_{j+2})}{P[\rho_j,\rho_{j+1})} = C^{1-\alpha}. \qquad (54)$$

Daraus ergibt sich: Wenn die Längen der Häufigkeitsteilintervalle eine wachsende geometrische Folge mit der Basis C bilden, dann bilden die Umfänge der betreffenden Häufigkeitsklassen eine abnehmende geometrische Folge mit der Basis $C^{-\alpha}$ und die Gewichte dieser Klassen eine Folge mit der Basis $C^{1-\alpha}$. D.h., wenn man die Funktion $v(\rho_j,\rho_{j+1})$ in ein Koordinatensystem mit doppellogarithmischem Maßstab gegen ρ_j auf der Abszisse einträgt, so erhält man eine Gerade mit dem Steigungskoeffizienten $-\alpha$. Auch für die Gewichte der Häufigkeitsklassen erhält man eine Gerade, jedoch mit dem Steigungskoeffizienten $1-\alpha$.

Wenn also die Rangordnungsdarstellung der Häufigkeitsstruktur laut (1) im allgemeinen im Bereich der großen Häufigkeiten eine durch den Parameter B erfaßte Krümmung hat, so erscheint die Darstellung dieser Struktur mit Hilfe der Umfänge oder der Gewichte der geometrischen Häufigkeitsklassen im doppellogarithmischen Maßstab immer als eine Gerade. Wegen dieses Umstandes ist die Zerlegung der Menge der Häufigkeiten in Häufigkeitsklassen ein effektives Verfahren zum Test, ob das Zipf-Mandelbrotsche Gesetz in seiner kanonischen Form erfüllt ist, sowie zur Schätzung des Parameters α. Im Falle kleiner Stichproben empfiehlt es sich, die Zerlegung in Teilintervalle auf der Achse der absoluten (nicht relativen) Häufigkeiten durchzuführen, indem man mit 1 anfängt und die Konstante C gleich 2 setzt[4]. Da die größte beobachtete Häufigkeit im allgemeinen nicht zweiter Ordnung ist, braucht man bei der Schätzung von α aus dem Diagramm das letzte Teilintervall nicht zu berücksichtigen[5].

Beispiele der Darstellung des empirischen Materials in Form von Häufigkeitsklassen wurden in Orlov (1970) gebracht; alle dort untersuchten Strukturen sind durch $\alpha = 1$ charakterisiert. Auf der Abb. 5 findet man die Diagramme von Umfängen der Häufigkeitsklassen über Verknüpfungen von je zwei (A), je drei (B), je vier (C) und je fünf (D) aufeinanderfolgenden melodischen Intervallen in Nocturnen von F. Chopin [die Berechnung wurde in Orlov (1970) erklärt]. Es wurde insbesondere untersucht, wie die Größe α beim Übergang zu einer Einheit höheren Typs zunimmt[6].

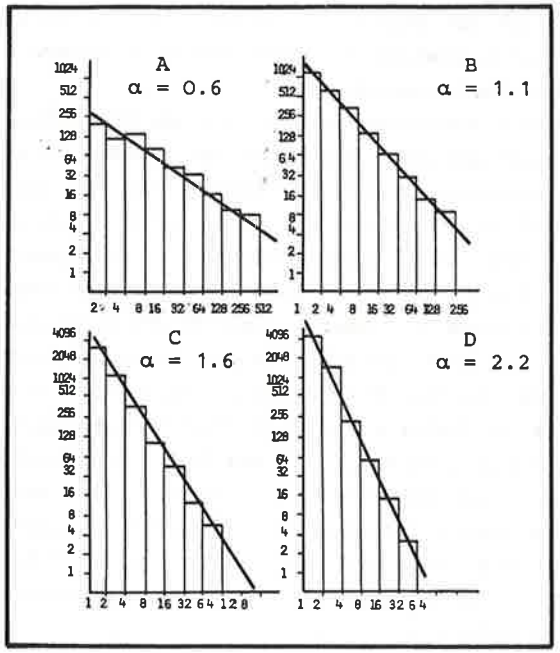


Abb. 5. Häufigkeitsstrukturen der Folgen melodischer Intervalle in Chopins Nokturnen (eigene Zählungen zusammen mit M.G. Boroda)

Anmerkungen

1 Es ist interessant, diesen Ausdruck mit der Häufigkeitsreihe $p_i = pe^{Bi}$, die Mandelbrot (1957) erhalten hat, zu vergleichen. In Mandelbrots Modell stellt diese Reihe die Häufigkeitsstruktur einer degenerierten "Sprache" dar, deren Alphabet aus einem Symbol und dem Zeichen für die Pause besteht.

2 Die Methoden der approximativen Lösung dieser Gleichung wurden in Orlov (1978a, 1980) dargelegt.

3 Zur Zeit der Entstehung dieser Arbeit (1973-1974) analysierte der Verfasser eine große Menge von empirischem Material auf verschiedenen linguistischen Ebenen. Es hat sich gezeigt, daß in den meisten Fällen die Wahl des Parameters Z völlig ausreichte; bisweilen mußte man auch den Wert von p_1 wählen (z.B. in georgischen Texten übersteigt die Häufigkeit des häufigsten Wortes 3-5 mal die Häufigkeit des zweithäufigsten Wortes. In solchen Fällen führt der Vergleich von p_1 mit der beobachteten Häufigkeit des häufigsten Wortes zu einer schlechten Leistung des Modells, und man muß p_1 so wählen, daß die Summe der Häufigkeiten der ersten 20-30 häufigsten Wörter im Modell der entsprechenden Summe im Text gleich ist). Bei der Wahl von α ergaben sich jedoch keine Schwierigkeiten und daher ist es plausibel anzunehmen, daß α = 1 der universelle Wert dieser Konstanten ist, unabhängig von der Art linguistischer Einheiten (außer Buchstaben und anderen Einheiten niedrigerer Ebenen).

4 Solche Häufigkeitsklassen nennt man in der Literatur gewöhnlich Oktaven. Ein Beispiel der Zerlegung in Oktaven findet man in Nadarejšvili & Orlov (1969).

5 Der folgende Fall ist z.B. gut möglich: die Häufigkeit des häufigsten Wortes im Text ist gleich 1050 und sie fällt in die Oktave 1024-2048; daher ist der größte Teil dieser Oktave "leer". Dadurch wird der reguläre Verlauf der "Treppen" (wie in Abb. 5) in dieser letzten Oktave gestört.

6 Die Zählung der Folgen von melodischen Intervallen erfolgt durch die sogenannte "gleitende Erhebung". Numeriert man alle Intervalle im Text laufend, so untersucht man zuerst die Folge mit den Zahlen 1,2,3, danach die mit 2,3,4, dann mit 3,4,5 usw. analog wie man es mit Digrammen, Trigrammen u.a. in linguistischen Massiven tut.

LITERATUR

ARAPOV, M.V., EFIMOVA, E.N., ŠREJDER, Ju.A., O smysle rangovych raspredelenij [Über den Sinn von Rangverteilungen]. Naučnotechniceskaja informacija Ser. 2, 1975, (Nr. 1), 9-20

BOOTH, A.D., A "law" of occurrence for words of low frequency. Information and Control 10, 1967, 409-418

BORODA, M.G., O častotnoj strukture muzykal'nych soobščenij [Über die Häufigkeitsstruktur musikalischer Nachrichten]. Soobščenija AN GSSR 76, 1974 (Nr. 2), 173-202

FRUMKINA, R.M., K voprosu o tak nazyvaemon zakone Cipfa [Zur Frage des sogenannten Zipfschen Gesetzes]. Voprosy jazykoznanija 1961, Nr. 2, 117-122

GAČEČILADZE, T.G., ELIAŠVILI, A.I., Statistika bukv sovremennogo literaturnogo gruzinskogo jazyka [Buchstabenstatistik der modernen georgischen Literatursprache]. Soobščenija AN GSSR 20, 1958 (Nr. 5), 565-567

GUIRAUD, P., Problèmes et méthodes de la statistique linguistique. Paris, Presses universitaires de France 1960

JOOS, M., Rezension von Zipf (1935). Language 12, 1936, 196-210

KALININ, V.M., Funkcionaly, svjazannye s raspredeleniem Puassona, i statističeskaja struktura teksta [Die mit der Poisson-Verteilung zusammenhängenden Funktionale und die statistische Struktur des Textes]. Trudy matematičeskogo institute imeni V.A. Steklova 29, 1965, 182-197

LEBEDEV, D.S., GARMAŠ, V.A., Statističeskij analiz trechbukvennych soćetanij russkogo teksta. Problemy peredači informacii 2, 1958, 78-80

MANDELBROT, B., O rekurrentnom kodirovanii, ograničivajuščem vlijanie pomech [On recurrent noise limiting coding]. In: Siforov, V.I. (Hrsg.), Teorija peredači soobščenij. Moskva, Izdatel'stvo inostrannoj literatury 1957, 139-157

NADAREJŠVILI, I.S., ORLOV, Ju.K., Ob upotreblenii slov različnoj častnosti v poeme Rustaveli "Vitjaz' v tigrovoj škure" [Über die Verwendung von Wörtern unterschiedlicher Häufigkeit in Rustavelis Poem "Der Held im Tigerfell"]. Soobščenija AN GSSR 55, 1969 (Nr. 2), 505-508

NADAREJŠVILI, I.S., ORLOV, Ju.K., Rost leksiki kak funkcija dliny teksta (na primere proizvedenij L.N. Tolstogo i Dž. Džojsa) [Das Anwachsen der Lexik als Funktion der Textlänge (am Beispiel der Werke von L.N. Tolstoj und J. Joyce)]. Soobščenija AN GSSR 64, 1971 (Nr. 3), 549-552

ORLOV, Ju.K., Obobščenie zakona Cipfa [Eine Verallgemeinerung des Zipfschen Gesetzes]. In: Rozencvejg, V.Ju. (Hrsg.), Problemy prikladnoj lingvistiki. Moskva, MGPIIJa 1969a, 255-261

ORLOV, Ju.K., Leksičeskij spektr literaturnych tekstov i dinamika rosta slovarja [Das lexikalische Spektrum literarischer Texte und die Dynamik des Vokabularzuwachses]. In: Rozencvejg, V.Ju. (Hrsg.), Problemy prikladnoj lingvistiki. Moskva, MGPIIJa 1969b, 250-254

ORLOV, Ju.K., Obobščenie zakona Cipfa-Mandel'brota [Eine Verallgemeinerung des Zipf-Mandelbrotschen Gesetzes]. Soobščenija AN GSSR 57, 1970 (Nr. 1), 37-40

ORLOV, Ju.K., Častotnye struktury konečnych soobščenij v nekotorych estestvennych informacionnych sistemach [Die Häufigkeitsstrukturen endlicher Nachrichten in einigen natürlichen Informationssystemen]. Habilitationsschrift, Tbilisi 1974

ORLOV, Ju.K., Obobščennyj zakon Cipfa-Mandel'brota i častotnye struktury informacionnych edinic različnych urovnej [Das verallgemeinerte Zipf-Mandelbrotsche Gesetz und die Häufigkeitsstrukturen der Informationseinheiten verschiedener Ebenen]. In: Guseva, E.K. (Hrsg.), Vyčislitel'naja lingvistika. Moskva, Nauka 1976, 179-202

ORLOV, Ju.K., Model' častotnoj struktury leksiki [Das Modell der Häufigkeitsstruktur der Lexik]. In: Andrjuščenko, V.M. (Hrsg.), Issledovanija v oblasti vyčislitel'noj lingvistiki i lingvostatistiki. Moskva, Moskovskij Gosudarstvennyj Universitet 1978 , 59-118

ORLOV, Ju.K., Informacionnye potoki: statističeskij analiz i prognozirovanie [Informationsflüsse: Statistische Analyse und Prognostizierung]. Naučno-techničeskaja informacija, Serija 2, 1980 (Nr. 2), 23-30

SIMON, H.A., Some further notes on a class of skew distribution functions. Information and Control 3, 1960, 90-98

VOLOŠIN, B.A., ORLOV, Ju.K., Obobščennyj zakon Cipfa-Mandel'brota i raspredelenie dolej cvetovych ploščadej v proizvedenijach živopisi [Das verallgemeinerte Zipf-Mandelbrotsche Gesetz und die Verteilung der Farbflächen in Gemälden]. Tbilisi, Institut kibernetiki AN GSSR 1972

YULE, G.U., The statistical study of literary vocabulary. London, Cambridge University Press 1944

ZIPF, G.K., Human behavior and the principle of least effort. Cambridge (Mass.), Addison-Wesley Press 1949

Ein Modell der Häufigkeitsstruktur des Vokabulars

Ju.K. Orlov

In den letzten 50 Jahren sind auf dem Gebiet der Lexikostatistik zahlreiche experimentelle Arbeiten (Häufigkeitswörterbücher, Konkordanzen, usw.) wie auch theoretische Arbeiten (vgl. Zipf 1935, 1949; Mandelbrot 1953, 1957 und viele andere) erschienen, die der Untersuchung und der theoretischen Begründung bzw. Verallgemeinerung empirischer Gesetzmäßigkeiten gewidmet waren. Im Augenblick befindet sich jedoch die Untersuchung der quantitativen Seite des Problems in einer gewissen Krise. Auch wenn man die Häufigkeitswörterbücher für die Kompilation von Minimalwörterbüchern, für maschinelle Übersetzung usw. verwenden kann, so fehlen jegliche verallgemeinernde Arbeiten über die Statistik des Lexikons.

Der Grund liegt in einer gewissen Kluft zwischen der Intention der theoretischen Arbeiten, die üblicherweise das Zipfsche Gesetz irgendwie "begründen" möchten, und den aktuellen Bedürfnissen der linguostatistischen Praxis.

Die allgemeine Kenntnis des Umstandes, daß der Bereich der Wörter mit mittlerer Häufigkeit dem Zipfschen Gesetz folgt (vgl. Frumkina 1961; Finkenstaedt & Wolf 1969), erlaubt es nicht, irgendwelche Rechnungen durchzuführen; ganz unbefriedigend ist das sogenannte Maß der lexikalischen Konzentration, das einen Vergleich des relativen Vokabularreichtums in Texten unterschiedlicher Länge erlauben sollte. Die meistbenutzte Beziehung R = = v/N (v = Stichprobenvokabular, N = Stichprobenumfang) ist ganz ungeeignet, da sich dieses Verhältnis innerhalb eines Textes beträchtlich ändert. Beispielsweise in Puškins "Kapitänstochter" (nach Angaben von Josselson 1953) beträgt es in einer Stichprobe von 5000 Wortwendungen 0,313; bei 10000 Wortwendungen 0,243 und im gesamten Text (N = 29345) 0,163. Unseres Erachtens sind jegliche Formeln, die diese Beziehung enthalten, ungeeignet (vgl.

Tešitelová 1972). Etwas besser ist der Vorschlag von Guiraud (1954): $R = v/\sqrt{N}$; aber auch diese Größe kann sich im Rahmen eines Textes ändern (für die "Kapitänstochter" ergeben sich die entsprechenden Zahlen: 22,2; 24,3; 27,9).

Wenn man nur das Vokabular bei <u>einem</u> Textumfang kennt, so kann man es für einen anderen Umfang nicht "umrechnen"; sollte diese Möglichkeit gefunden werden, so würde sich das o.a. Problem von alleine lösen. Manchmal werden für empirische Daten die Werte der Koeffizienten K, B und γ der Mandelbrotschen Formel

$$p_i = \frac{K}{(B+i)^{\gamma}}, \qquad (1)$$

(wo p_i die Häufigkeit des i-ten Wortes in einem nach abnehmender Häufigkeit geordneten Wortinventars ist) berechnet, aber die linguistische Bedeutung dieser Konstanten ist unklar, und es ist schwer, irgendwelche Gesetzmäßigkeiten ihrer Veränderung beim Übergang von einem Text zum anderen festzustellen. Die Menge dieser Konstanten, die die Häufigkeitsstruktur einer konkreten Stichprobe beschreibt, ist ebensoweinig mit der analogen Menge einer anderen Stichprobe vergleichbar wie die Häufigkeitsreihen selbst. Deswegen entstand schon längst die Notwendigkeit, ein mathematisches Modell der Häufigkeitsstruktur des Lexikons aufzustellen, das die Lösung der aufgezählten Probleme und darüberhinaus weiterer Probleme ermöglicht.

In der vorliegenden Arbeit wird ein Modell dargestellt, das folgendes leistet:

(1) Die Berechnung des Vokabulars, der Häufigkeitskurve und des lexikalischen Spektrums (Zahl der einmaligen, zweimaligen usw. Wörter) bei beliebigem Testumfang, wenn das Vokabular in einer Stichprobe aus dem gegebenen Text bekannt ist.

(2) Schätzung der möglichen zufälligen Abweichung des tatsächlichen Vokabulars von der theoretischen Voraussage.

(3) Schätzung des Grads der "statistischen Ähnlichkeit"zweier oder mehrerer Texte, die in eine Stichprobe zusammengefaßt wurden.

(4) Charakterisierung des Vokabularwachstums in einem gegebenen Text und des lexikalischen Spektrums in einem beliebigen Abschnitt durch einen den Daten des Texts (oder der Stichprobe) angepaßten Parameter, der direkt als eine Konvention zur Messung des relativen Vokabularreichtums verwendet werden kann.

1.

Bei der Ableitung des Modells gehen wir von einigen Resultaten von Kalinin (1964, 1965) aus. Im folgenden stellen wir kurz diejenigen Ergebnisse dar, die wir in der vorliegenden Arbeit verwenden werden.

Kalinin analysiert den "unbegrenzten Text", in dem jedes Wort eine a priori Vorkommenswahrscheinlichkeit hat, die von der Position im Text und von den früher verwendeten Wörtern unabhängig ist. Obwohl diese Voraussetzung streng genommen in den reellen Texten nicht erfüllt wird, erlaubt sie trotzdem, einige Beziehungen abzuleiten, die mit der Realität annehmbar übereinstimmen.

Die Unabhängigkeitsannahme erlaubt es, die Beziehung zwischen der mathematischen Erwartung des Vokabularumfanges $v(N)$ und der Zahl der m-mal vorkommenden Wörter $v_m(N)$ bei verschiedenen Umfängen N abzuleiten.

Wenn die mathematischen Erwartungen $E[v(N_o)]$ und $E[v_m(N_o)]$ bei dem Umfang N_o bekannt sind, so gilt bei beliebigem Umfang N

$$E[v(N)] = E[v(N_o)] - \sum_{j\geq 1} E[v_j(N_o)]\left(1 - \frac{N}{N_o}\right)^j \qquad (2)$$

und

$$E[v_m(N)] = \frac{1}{m!}\left(\frac{N}{N_o}\right)^m \sum_{j\geq 1} E[v_j(N_o)]j(j-1)\ldots(j-m+1)\cdot$$

$$\cdot \left(1 - \frac{N}{N_o}\right)^{j-m}, \qquad (3)$$

wo E die mathematische Erwartung bedeutet.

Aus diesen Beziehungen sind die uns unbekannten Wahrscheinlichkeiten des Wortvorkommens ausgeschlossen; dabei ist die Umrechnung sowohl "rückwärts" (für $N < N_o$) als auch "vorwärts" (für $N > N_o$) möglich. (Wenn man anstelle der mathematischen Erwartungen $E[v(N_o)]$ und $E[v_m(N_o)]$ nur die beobachteten Werte $\hat{v}(N_o)$ und $\hat{v}_m(n,N_o)$ beim Umfang N_o verwendet, dann ist nur die Umrechnung "rückwärts" möglich.) Für die Varianz der Größen $v(N)$ und $v_m(N)$ gibt Kalinin folgende Beziehungen an:

$$V[v(N)] = E[v(2N)] - E[v(N)] - \frac{\{[E[v_1(N)]\}^2}{N} + \frac{E[v_2(2N)]-E[v_2(N)]}{N} + \varepsilon \qquad (4)$$

$$V[v_m(N)] = E[v_m(N)] - \frac{E[v_{2m}(2N)]}{\sqrt{\Pi m}} - \frac{\{E[v_m(N)]\}^2}{N} + \varepsilon \qquad (5)$$

wo V die Varianz bedeutet.

Wenn man also das erwartete lexikalische Spektrum des Textes, beschrieben durch $E[v_m(N_o)]$, bei einem gegebenen Umfang N_o kennt, dann kann man die Häufigkeitsstrukturen des Textes bei einem beliebigen anderen Umfang berechnen und die Zufälligkeit bzw. Signifikanz der beobachteten Abweichungen beurteilen.

In dem dargelegten Modell spielt eine grundlegende Rolle der Umfang Z, der die Häufigkeitsstrukturen nach dem Zipf-Mandelbrotschen Gesetz gestaltet (vgl. Orlov 1970). Die formale Analyse einer derartigen Struktur, wie in Orlov (1976) durchgeführt, beruht auf folgenden Annahmen:

(a) Das Vorkommen von hapax legomena (einmal vorkommende Wörter) im Text ist obligatorisch (eine Beobachtung von Yule 1944);

(b) die Größe γ in der Mandelbrotschen Formel (1) ist gleich eins.[1)]

Dies führt zu folgenden Beziehungen:

$$p_i = \frac{\frac{1}{\ln (Zp_{max})}}{\frac{1}{p_{max} \ln (Zp_{max})} - 1 + i}; \qquad (6)$$

$$v(Z) = \frac{Z - \frac{1}{p_{max}}}{\ln (Zp_{max})}; \qquad (7)$$

$$v_m(Z) = \frac{v(Z)}{m(m+1)}, \qquad (8)$$

wo p_{max} die größte im Text gefundene relative Worthäufigkeit ist.

Vergleicht man (6) mit (1), so sieht man, daß sie identisch sind, wenn

$$K = \frac{1}{\ln Zp_{max})}; \quad B = \frac{K}{p_{max}} - 1; \quad \gamma = 1. \qquad (9)$$

Die Werte von K und B kann man also mit Hilfe von Z und P_{max} darstellen.

Betrachtet man die durch (8) bestimmten Größen $v_m(Z)$ als die mathematischen Erwartungen der Anzahl unterschiedlicher m-mal vorkommender Wörter beim Textumfang Z[2)], dann kann man durch Einsetzung von (7) und (8) in (2) und (3) den Vokabularumfang v(N,Z) und die Anzahl $v_m(N,Z)$ m-maliger Wörter bei beliebigem Umfang N erhalten (hier lassen wir das Zeichen der mathematischen Erwartung weg):

$$v(N,Z) = v(Z) \sum_{j=0}^{\infty} \frac{(1-X)^j}{j+1}; \qquad (10a)$$

$$v_m(N,Z) = v(Z) \sum_{j=0}^{\infty} \frac{(1-X)^j}{(j+m)(j+m-1)}, \qquad (11a)$$

wo $X = Z/N$[3)]. Wenn X von 1 stark abweicht, dann bekommen diese Größen eine andere Form.[4)]

$$v(N,Z) = v(Z) \frac{\ln X}{X - 1}; \qquad (10b)$$

$$v_1(N,Z) = \frac{v(Z) - Xv(N,Z)}{1 - X}; \qquad (11b)$$

$$v_m(N,Z) = \frac{v_{m-1}(N,Z) - v_{m-1}(Z)}{1 - X}; \quad m = 2,3,... \qquad (11c)$$

(Die Rechnung mit der Rekursionsformel (11c) muß auf mehrere Dezimalstellen durchgeführt werden, z.B. mit einem Mikrorechner; falls $X \approx 1$, so soll man lieber Formel (11a) benutzen.)[5)]

Unser Modell der Häufigkeitsstruktur des Lexikons stellt also einen unbegrenzten, statistisch homogenen Text dar, der folgende Eigenschaften hat:

(1) Er besitzt immer hapax legomena, d.h. das Vokabular wächst unbegrenzt mit dem Anwachsen des Textumfangs.

(2) In Stichproben mit einem festen Umfang Z ist die mathematische Erwartung der Anzahl m-maliger Wörter und die des Vokabularumfangs durch (7) und (8) und die geordnete Häufigkeitsreihe durch (6) gegeben.

(3) In Stichproben mit beliebigem Umfang N ist die mathematische Erwartung der Anzahl m-maliger Wörter durch (11a) und (11b) und die des Vokabularumfangs durch (10a) und (10b) gegeben. Den letzten Ausdruck kann man als die Funktion des Vokabularanwachsens betrachten, die mit der Zunahme des Textumfangs in Zusammenhang steht, wenn Z

die Rolle eines Parameters hat. Der Verlauf der Häufigkeitskurve im Bereich der häufigen Wörter (deren relative Häufigkeiten sich bei der Änderung des Stichprobenumfangs nicht ändern) wird auch in diesem Fall durch (6) beschrieben.

Mit anderen Worten, wenn ein Text mit Umfang Z dem Zipf-Mandelbrotschen Gesetz folgt, dann ist dieser Umfang der einzige, bei dem das Gesetz erfüllt werden kann. Bei Stichproben mit einem beliebigen anderen Umfang sind Abweichungen unvermeidlich. Unten werden wir zeigen, daß die bekannte "Abweichung des Schweifes" im Bereich seltener Wörter eine Folge gerade dieses Umstandes ist. Den Umfang Z werden wir im weiteren als den "Zipfschen Umfang" bezeichnen.

Diese Größe Z ist eine wichtige Charakteristik des Textes. Außer der Möglichkeit, das Vokabular und das Spektrum bei beliebigem Umfang zu berechnen, liefert sie unmittelbar ein geeignetes Maß der lexikalischen Konzentration, ein Maß des relativen Vokabularreichtums. Wenn es also zwei Texte mit unterschiedlichen Werten von Z gibt so, daß $Z_1 > Z_2$, dann ist bei gleichem p_{max}

$$v(N,Z_1) > v(N,Z_2) \qquad (12)$$

für beliebiges N. Das heißt, eine Stichprobe mit Umfang N aus dem ersten Text wird ein reicheres Vokabular haben als eine Stichprobe desselben Umfangs aus dem zweiten Text, dessen Zipfscher Umfang kleiner ist. Mit anderen Worten, wenn man die Zipfschen Umfänge zweier Texte vergleicht, dann wird der Text ein reicherer Vokabular haben, dessen Z größer ist. Dies erlaubt uns, die Größe Z als ein konventionelles Maß des relativen Vokabularreichtums zu nennen.

Damit ist die Aufstellung des Modells beendet. Alles weitere hängt davon ab, wie man dieses Modell verwendet. In Orlov (1970, 1976) wurde festgestellt, daß Z im Extremfall der Länge eines großen literarischen Werkes entspricht.[6] Würden wir nämlich die volle Länge solcher Texte als Z in (6), (7) und (8) einsetzen, so erhielten wir eine Übereinstimmung der beobachteten und der

theoretischen Werte innerhalb der ± 20% Toleranz für alle Werke, deren Länge über 10-20 Tausend Wortverwendungen betrug. Stichproben aus diesen Texten würden mit ungefähr derselben Genauigkeit den Ausdrücken (10) und (11) (vgl. Orlov 1969) folgen. Es hat sich gezeigt, daß man in sehr großen Texten eine bessere Übereinstimmung erreichen kann, wenn man als Z die Umfänge der vom Autor festgelegten Teile dieser Texte (Bände, Bücher u.a.), einsetzt (vgl. Nadarejšvili, Orlov 1971), obwohl in einigen Fällen (z.B. in Joyce "Ulysses") die bessere Übereinstimmung bei vollem Text erfolgte.

Wie interessant diese Erscheinung an sich auch sein mag (der relative Vokabularreichtum wird durch die Textlänge bestimmt: die Wachstumskurve des Vokabulars ist desto steiler, je mehr der Autor zu schreiben vorhat) ihre Überprüfung würde durch die im allgemeinen niedrige Genauigkeit der theoretischen Prognose erschwert. Die Uneindeutigkeit der Prognose für große Texte (der Zipfsche Umfang kann sowohl der vollen Textlänge als auch der Länge eines Teiles entsprechen), die Unmöglichkeit der Prognose für kurze Texte (kürzer als 10000 Wortverwendungen) und die Unklarheit der Grenzen der "Zwischenzone" (10000-?), in der einige Texte mit ihrem Umfang den Gesetzmäßigkeiten (6) - (8) folgen und andere wiederum nicht, entwertet diese Hypothese (Z = volle Textlänge) für praktische Rechnungen.

Es wurde jedoch beobachtet, daß die Häufigkeitsstruktur kurzer Texte der Struktur von Abschnitten aus längeren Texten ähnelt, die in ihrer vollen Länge dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgen und dieselbe charakteristische Krümmung der Treppenkurve oberhalb der theoretischen Kurve (6) im Bereich seltener Wörter haben. Diese Beobachtung führte zu der Hypothese: <u>für jeden Text gibt es einen eigenen Zipfschen Umfang</u> unabhängig sowohl von seiner Häufigkeitskurve als auch von seinem reellen Umfang. Mit anderen Worten, entweder kann man aus beliebigem Text Stichproben vom Umfang Z erheben, deren Vokabular und Häufigkeitsstruktur bis auf zufällige Abweichungen durch (6) - (8) genau beschrieben werden, oder man kann den Text (wenn sein Umfang kleiner als Z ist) als einen Abschnitt aus einem Text mit Umfang Z, für den (6) - (8) gelten, betrachten. Wenn

man Z kennt, so kann man in beiden Fällen das Vokabular und das Häufigkeitsspektrum bei beliebigem Umfang aufgrund von (10) und (11) berechnen.

Den letzten Umstand benutzen wir zur Überprüfung der aufgestellten Hypothese. Hat man den Wert von Z für den gegebenen Text geschätzt, dann berechnet man die theoretischen Werte derjenigen Textparameter, die zur Bestimmung von Z nicht direkt verwendet werden. Wenn die aufgestellte Hypothese korrekt ist, dann muß zwischen dem Vokabular und dem lexikalischen Spektrum des Textes bei Stichproben mit beliebigem Umfang der durch (10a), (10b) und (11b) gegebene Zusammenhang bestehen.

Um den Zipfschen Umfang zu schätzen, löst man bezüglich Z die Gleichung

$$v(N,Z) = \hat{v}(N) \qquad (13)$$

wo $\hat{v}(N)$ das beobachtete Vokabular der Stichprobe aus dem Text (eventuell der ganze Text), der den Umfang N hat, darstellt. Mit anderen Worten, wir führen die theoretische Kurve der Vokabularzunahme durch einen einzigen Punkt, nämlich durch den beobachteten Wert des Wortschatzes.

Der Ausdruck (13) ist bezüglich Z leider transzendental und läßt sich mit Hilfe einfacher Funktionen nicht ausdrücken.[7)]

Wenn das Vokabular $\hat{v}^{(')} = v(N_1,Z)$ beim Umfang N_1 gegeben ist, dann kann man offensichtlich nach der Bestimmung von Z für das Vokabular $v(N_2,Z)$ einen Umfang N_2 berechnen. Schätzen wir nun mit Hilfe unseres Modells den Grad möglicher zufälliger Abweichungen des beobachteten Vokabulars von dem prognostizierten beim Umfang N_2.

Wenn man Z genau kennen würde, dann könnte man die Dispersion der Abweichungen des beobachteten Wertes von dem prognostizierten Wert direkt mit Kalinins Formel (4) bestimmen. Setzt man in sie den Ausdruck für den Wortschatz (10b) ein, und behält man nur die ersten zwei Glieder der Entwicklung, so kann man schreiben:

$$V[v(N,Z)] = v(Z)\left(\frac{\ln\frac{Z}{2N}}{\frac{Z}{2N}-1} - \frac{\ln\frac{Z}{N}}{\frac{Z}{N}-1}\right) =$$

$$= v(N,Z)\left(\frac{\ln\frac{X}{2}}{\ln X}\cdot\frac{X-1}{\frac{X}{2}-1} - 1\right) = v(N,Z)\cdot G(X), \qquad (14)$$

wo

$$X = \frac{Z}{N};\ G(X) = \left(1 - \frac{\ln 2}{\ln X}\right)\cdot\frac{X-1}{\frac{X}{2}-1} - 1 \text{ ist.}$$

Für praktische Rechnungen ist es angebracht, die relative Standardabweichung in Prozenten zu benennen. Sie gleicht der Standardabweichung, dividiert durch den Erwartungswert und multipliziert mit 100, d.h.

$$\delta = \frac{\sqrt{V[v(N,Z)]}}{v(N,Z)}\cdot 100\% = \sqrt{\frac{G(x)}{v(N;Z)}}\cdot 100\%.$$

Der genaue Wert Z ist aber unbekannt, und wir schätzen ihn lediglich aus den Daten der Stichprobe. Der Fehler der Prognose setzt sich zusammen aus der zufälligen Streuung in dem prognostizierten Punkt und der Verschiebung des prognostizierten Punktes selbst durch zufällige Abweichungen des Wortschatzes im beobachteten Punkt.

Die relative Varianz der Prognosefehler kann man als die Summe

$$\delta^2 = \delta_1^2 w^2 + \delta_2^2 \qquad (15)$$

darstellen, wo

$$\delta_1 = \sqrt{\frac{G(X_1)}{v(N_1,Z)}} \cdot 100\%; \quad \delta_2 = \sqrt{\frac{G(X_2)}{v(N_2,Z)}} \cdot 100\%;$$

$$X_1 = \frac{Z}{N_1}\ ; \quad X_2 = \frac{Z}{N_2}\ ; \quad v(N,Z) = \hat{v}^{(1)}$$

ist.

Der Koeffizient W berücksichtigt die geometrischen Eigenschaften der Verschiebung des prognostizierten Punktes infolge der Verschiebungen des beobachteten Punktes.[8] Ein exakter analytischer Ausdruck ist aus demselben Grund nicht möglich wie die Lösung der Gleichung (13); eine annehmbare Approximation ist gegeben durch

$$W \approx \left(\frac{\lg v(N_2,Z)}{\lg \hat{v}^{(1)}}\right)^4 . \qquad (16)$$

Im Vergleich mit der exakten Berechnung der Veränderung des Abstandes zwischen den Kurven der Familie (10) bei unterschiedlichen Z ergibt sie einen relativen Fehler, der unter 6-8% liegt.

Auf diese Weise können wir also den Zufälligkeitsgrad der festgestellten Unterschiede zwischen der Voraussage $v(N_2,Z)$ und dem tatsächlich beobachteten Vokabularumfang beim Umfang N_2 abschätzen. Wir werden schließen, daß die festgestellte relative Abweichung rein zufällig ist, wenn sie 3δ nicht überschreitet. Dies entspricht einer Wahrscheinlichkeit größer als 0.003.

Wir bemerken, daß diese Schätzung für solche Stichproben gilt, die unabhängig aus dem Modelltext erhoben wurden. Wenn die Stichprobe mit dem Umfang N_1, für die man eine Prognose aufstellt, ein Teil der Stichprobe mit Umfang N_2 ist (oder umgekehrt) und dieser Umfang nur unbedeutend größer (kleiner) als N_2 ist, so vergrößert sich wesentlich die Genauigkeit der Prognose. Eine derartige Erhöhung der Prognosegenauigkeit wird in den Tabellen im Anhang demonstriert, wenn sich der Stichprobenumfang dem Punkt nähert, für welchen die Prognose aufgestellt

wurde. Falls der Umfang der Unterstichprobe viel kleiner ist als der der großen Stichprobe, so kann man bei diesen Prognosen beide Stichproben als unabhängig betrachten.

2.

Das Material, an dem die Adäquatheit des dargelegten Modells überprüft wurde, ist in den Tabellen 1a, 1b, 2a und 2b (s. Anhang 2) aufgeführt. Die benutzten Werke wurden durchlaufend numeriert, damit wir auf sie mit einer Zahl hinweisen können. Die Hinweiszahl wird in geschweifte Klammern gesetzt. Bevor wir die Tabellen analysieren, demonstrieren wir an einigen Beispielen ausführlich die Übereinstimmung des dargelegten Modells mit konkreten lexikalischen Stichproben.

In der Abb. 1 werden die Häufigkeitskurven des Romans "Kapitänstochter" von Puškin {28}[9] (obere Kurve) und ein Abschnitt aus dem Roman im Umfang von 5000 Wortverwendungen {30} dargestellt. Die Punkte und die dicken Striche stellen die beobachteten Häufigkeiten dar. Die halbdicken ununterbrochenen Kurven sind die theoretischen Häufigkeiten, die aufgrund von (6), in unserem Fall für Z = 45000, berechnet wurden. Die theoretischen Werte der Anzahl seltener Wörter, berechnet nach (11) für den Umfang N = = 29345, sind als "Treppenkurve" mit dünnen Strichen aufgetragen. Man kann leicht sehen, daß in diesem Fall, d.h. wenn N dem Z relativ nah steht, eine gute Übereinstimmung zwischen der Häufigkeitsreihe und dem genannten Verlauf der theoretischen Kurve (6) erreicht wird. Auf der Graphik des Abschnitts sieht man, daß der Treppenabschnitt im Bereich seltener Wörter (d.h. sowohl die empirischen "Rechtecke" als auch die theoretischen "Treppen") überwiegend über der theoretischen Kurve liegt und die typische "Krümmung des Schweifes" aufweist. Die Zahl der hapax legomena beträgt in diesem Fall mehr als die Hälfte des Vokabulars.

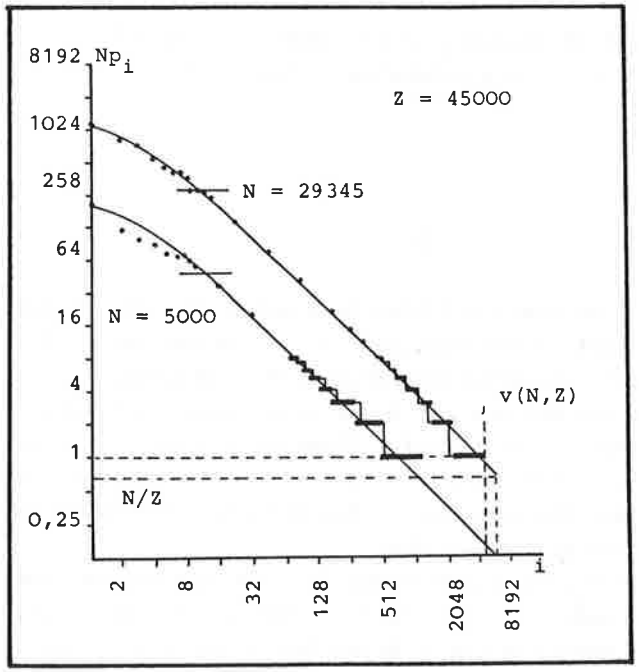


Abb. 1. Die Häufigkeitskurve von Puškins "Kapitänstochter" (obere Kurve) und von einem Abschnitt von 5000 Wortverwendungen {30}. In beiden Fällen ist laut (9) K = 0.134; B = 2.40. Die theoretische Kurve, berechnet nach (6) (die ununterbrochene halbdicke Kurve) läuft bis i = v(Z) = 6020, wo die Summe der relativen Häufigkeiten 1 erreicht, und $P_{v(Z)} = \frac{1}{Z}$.

Die obere Graphik auf der Abb. 2 stellt die Häufigkeitskurve der Gesamtwerke von Puškin {127} dar. Der Stichprobenumfang (N = 544777) überschreitet wesentlich den Zipfschen Umfang und die "Krümmung des Schweifes" erfolgt auf die andere Seite, d.h. unterhalb von (6). Obwohl in diesem Fall die Übereinstimmung der theoretischen "Treppenkurve" und der empirischen "Rechtecke" etwas schlechter ist[10], stimmt der allgemeine Krümmungstrend über-

ein. Die Zahl der hapax legomena ist hier kleiner als die Hälfte des Vokabulars.

Die mittlere Graphik der Abb. 2 stellt die Häufigkeitskurve der Novelle "Pique Dame" {25} dar. Die auffallende Ähnlichkeit dieser Kurve mit der des Abschnitts aus der "Kapitänstochter" (vgl. Abb. 1) ist die Konsequenz der Ähnlichkeit der grundlegenden Textparameter (N = 6861, Z = 35000) mit denen des Abschnitts {30}.

Auf der anderen Seite zeigt "Skazka o rybake i rybke" {5} (die unterste Graphik auf der Abb. 2), die einen erheblich kleineren Zipfschen Umfang hat (N = 948, Z = 2200), am Anfang einen viel langsameren Verlauf der Häufigkeitskurve (relativer Überfluß häufiger Wörter), wodurch der Einfluß von Z auf die lexikalische Konzentration erklärt wird - denn je mehr relativ häufige Wörter im Text vorhanden sind, desto weniger Platz bleibt für die anderen.

Die Abbildungen 1 und 2 zeigen anschaulich die Dynamik der Veränderung der lexikalischen Struktur mit der Veränderung des Stichprobenumfangs. Solange N < Z ist, liegt der treppenförmige Teil der Graphik oberhalb der theoretischen Häufigkeitskurve und nimmt langsamer ab; versucht man den Verlauf der ganzen Kurve mit der Mandelbrotschen Formel (1) zu approximieren, so ergibt sich, daß γ < 1 ist. Je mehr sich der Stichprobenumfang dem Zipfschen Umfang nähert, desto gerader wird die Graphik: die rechten Ecken der Treppen nähern sich der Häufigkeitskurve (6) und das tatsächliche Vokabular nähert sich der Zahl v(Z); die Zahl von hapax legomena erreicht fast die Hälfte des Vokabulars. Wenn der Stichprobenumfang den Zipfschen Umfang übersteigt, so biegt sich der treppenförmige Teil der Graphik nach unten.

Die Größe p_{max} hat einen geringen Einfluß auf den Verlauf der Häufigkeitskurve im Bereich der Wörter mit mittleren und niedrigeren Häufigkeiten. Es ist eben dieser Umstand, der es erlaubt, einen approximativen Wert von p_{max} zu verwenden, wenn sein genauer Wert für die gegebene Stichprobe unbekannt ist. In den aufgeführten Tabellen wurden solche Orientierungswerte mit dem Zeichen ~ versehen. Sie wurden für die Sprache der Stichprobe als maßgebend betrachtet, wenn die Quelle keine genauen Daten

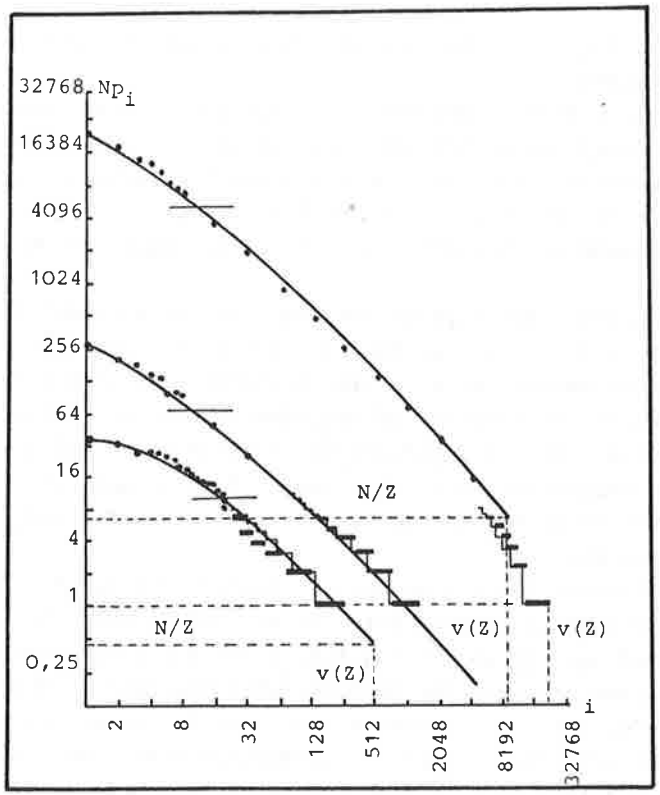


Abb. 2. Häufigkeitskurven der Gesamtwerke von A.S. Puškin {127} mit Z = 75000; K = 0,122; B = 1,54 (oben); "Pique Dame" {25} mit Z = 35000; K = 0,139; B = = 2,66 (in der Mitte) und "Skazka o rybake i rybke" mit Z = 2200; K = 0,228; B = 4,95 (unten). Wie hier und auf Abb. 1 ersichtlich, weisen die Häufigkeiten der häufigsten Wörter nichtreguläre Abweichungen von der Kurve (6) auf. Ähnliche nichtreguläre Abweichungen kann man auch an anderen Daten beobachten. Der Bereich, in dem diese Abweichungen aufhören, liegt bei den Häufigkeiten zwischen 0,01 und 0,005. Die relative Häufigkeit 0,01 ist auf den Abbildungen mit einem dünnen horizontalen Strich gekennzeichnet, der den oberen Teil jeder Graphik schneidet.

lieferte. Wenn uns für einige Sprachen überhaupt keine Angaben über den Wert von p_{max} zur Verfügung standen z.B. für die afghanischen Stichproben von Ludin {169 - 174}, dann haben wir p_{max} = 0,05 gesetzt und diesen Wert mit einem Fragezeichen versehen.

Die Abbildungen 1 und 2 haben lediglich eine illustrative Aufgabe. Alle Angaben über die Häufigkeitsstruktur des Lexikons, die uns zur Verfügung standen, sind in den Tabellen 1 und 2 enthalten. In Tab. 1a und 1b sind Angaben entweder über vollständige Texte einzelner Werke oder über Abschnitte aus ihnen enthalten. In Tab. 2a und 2b sind Daten von Vereinigungen mehrerer Texte zu einer Stichprobe aufgeführt.

Die Tabellen 1a und 2a enthalten Angaben über die Texte und Stichproben, deren Vokabulare sowie lexikalische Spektra bekannt waren. Dies ermöglichte, nach der Berechnung von Z die theoretische Anzahl der ein-, zwei- und dreimaligen Wörter mit den entsprechenden empirischen Werten zu vergleichen. Diese Größen wurden deswegen gewählt, weil sie, nach dem dargelegten Modell, am meisten vom Stichprobenumfang beeinflußt wurden und insgesamt etwa 3/4 des Wortschatzes bilden.

In den Tabellen 1b und 2b sind lediglich Angaben über Umfänge und Vokabulare einiger Texte und Stichproben zusammengestellt; daher mußten wir uns in diesen Tabellen auf die Berechnung des Zipfschen Umfangs beschränken (die hier mit einer geringeren Genauigkeit durchgeführt wurde; vgl. Anhang II).

Die Analyse der Tabellen 1a und 2a zeigt, daß die Übereinstimmung zwischen den theoretischen Prognosen und den beobachteten Werten in praktisch annehmbaren Grenzen liegt. Für alle Texte und Stichproben, deren Vokabular mehr als 1000 Wörter beträgt, wurde eine signifikante, im Durchschnitt dreiprozentige Unterbesetzung von hapax legomena bei dreimaligen Wörtern beobachtet. Bei zweimaligen Wörtern wurde im Schnitt keine signifikante Verschiebung der Prognose beobachtet.[11] Zwecks einer genaueren a priori Berechnung der Zahl von hapax legomena empfiehlt es sich, den Wert von (11) mit einem empirischen Korrekturkoeffizienten von 0,92 zu multiplizieren; für die dreimaligen Wörter mit 1,08. Nach der Durchführung dieser Korrekturen beträgt die relative

Standardabweichung der empirischen Größe von der Prognose für einmalige Wörter ± 10%, für zweimalige Wörter ± 11% und für dreimalige Wörter ± 15%. Eine Überprüfung mit dem χ^2-Kriterium zeigte, daß die Verteilung der Prognosefehler der Hypothese ihrer Normalität nicht widerspricht.

Diese Abweichungen, die üblicherweise größer als zufällig sind, stimmen gut mit der Streuung überein, die man in gleichlangen Stichproben aus einem Text vorfindet. Stichproben von jeweils 1000 Wortverwendungen aus Puškins "Eugen Onegin" (23 Stichproben), "Ruslan i Ljudmila" (11 Stichproben) und "Skazka o care Saltane" (4 Stichproben), die uns freundlicherweise A.Ju. Šajkevič zur Verfügung stellte, weisen folgende Standardabweichungen von ihren Mittelwerten (in %) auf:

für das Vokabular	± 6,43
für einmalige Wörter	± 9,85
für zweimalige Wörter	± 11,0
für dreimalige Wörter	± 20,8

Alle diese Abweichungen übersteigen die möglichen zufällig. Speziell die Standardabweichung des Vokabulars bei N = 1000 und Z = 240000 laut (14) ist gleich ± 3,6%. Der Chi-Quadrat-Test zeigte, daß die beobachteten Abweichungen eindeutig signifikant sind. Also gibt es in Texten nichtzufällige Schwankungen im Prozeß der Generierung neuer Wörter; ähnliche Erscheinungen kann man in einigen Daten in den Tabellen 3 und 4 beobachten.

In der Tabelle 3 sind Angaben über Texte enthalten, die aus mehr als einer Stichprobe bestehen (der volle Text und Abschnitte davon; verschiedene Teile eines Textes). Die Prognose für das Vokabular einer jeden Stichprobe erfolgt aus den anderen Stichproben desselben Textes. Mit N_1, v_1 und z_1 werden die Parameter der Stichprobe 1, für die man die Prognose macht, gekennzeichnet. Stichprobe 2 ist diejenige, aufgrund derer man die theoretische Prognose aufstellt.

Die Analyse der Abweichungen der empirischen Wortschatzumfänge von der Prognose zeigt keine wesentlichen Verschiebungen: die algebraische Summe aller mittleren Prognosefehler beträgt ungefähr 0,5%. Die relative Standardabweichung ist gleich ± 6,04%,

was mit der oben angegebenen Streuung des Wortschatzes in gleichlangen Abschnitten aus den Werken Puškins (± 6,43%) gut übereinstimmt. Es ist bemerkenswert, daß für die Texte, die nicht in Teile aufgeteilt sind, die Prognose genauer ist, als wenn man aufgrund eines Textteils die Prognose für einen anderen Textteil aufstellt. Während im ersten Fall die Abweichung die theoretische ± 3σ Grenze nicht oder nur geringfügig übersteigt, so findet man im zweiten Fall eine wesentlich größere Abweichung.

Solche signifikanten Abweichungen findet man auch bei Prognosen für die Texte der altgeorgischen Evangelien {6 - 10} (Prognosen aufgrund eines Evangeliums für ein anderes oder für den vollständigen Text), die in der Tabelle 3 nicht aufgeführt wurden. Diese Tatsache bezeugt eine gewisse "statistische Selbstständigkeit" der Teile eines großen Textes. Große Abweichungen gibt es auch bei Vereinigungen vieler Texte über Automobilbau {161 - 162} zu einer Stichprobe.

In den Tabellen 4a und 4b wird das Anwachsen des Vokabulars vom Anfang einiger Texte an im Detail untersucht. Als Ausgangsgröße wurde der Wortschatzumfang der ganzen untersuchten Stichprobe genommen. Es ist interessant, daß bei einigen Texten ("Slovo o polku Igoreve"; "Vojna i mir" und "Voskresenie" von Tolstoj) eine sehr genaue Übereinstimmung mit theoretischen Prognosen besteht (in Extremfällen keine Abweichung über 2-4%), während bei anderen ("Ilja Muromec i Kalin-car'"; "Krejcerova sonata" und "Kazaki" von Tolstoj) eine signifikante Herabsetzung (gekennzeichnet mit Sternchen) des tatsächlichen Wortschatzes beobachtbar ist, die im Vergleich mit dem theoretischen Wortschatz um 12 - 14% kleiner ist.[12)]

Fassen wir die Resultate des Vergleichs des dargelegten Modells mit empirischen Daten zusammen.

1. In dem untersuchten Material gab es keinen Text, dessen Häufigkeitsstruktur von dem Modell abweichen würde; mit anderen Worten, es besteht ein Zusammenhang zwischen dem Vokabularumfang in einem und der Häufigkeitsstruktur in einem beliebigen anderen Punkt des Textes. Die Häufigkeitsstruktur ist dynamisch und hängt von dem Textumfang ab. Der Parameter Z, berechnet aus den Angaben der Stichprobe, reicht sowohl für die Beschreibung des

Vokabulars innerhalb des Textes als auch für die Beschreibung der Häufigkeitsstruktur in einem beliebigen Textabschnitt.

2. Der Zipfsche Umfang Z scheint ein annehmbares Maß des relativen Vokabularreichtums zu sein. Obwohl die Streuung der Schätzungen von Z für unterschiedliche Stichproben aus einem Text auf den ersten Blick ziemlich groß ist (z.B. schwanken die Schätzungen für Puškins "Kapitänstochter" zwischen 33000 und 47000), ist die Streuung der "Kreuzprognosen" aufgrund dieser Schätzungen (Tab. 3) nicht signifikant groß.[13] Als Maß des relativen Vokabularreichtums kann man auch die Größe v(Z) benutzen; in dem Falle wird auch der Einfluß des Wertes von p_{max} berücksichtigt.

3. Die Analyse der Abweichungen empirischer Größen von ihren theoretischen Prognosen führt zu dem Schluß, daß bei zufriedenstellender Übereinstimmung, die im Durchschnitt besteht, einige Abweichungen als nicht zufällig zu betrachten sind. Dies kann entweder durch die Wirkung von Faktoren, die im Modell nicht berücksichtigt wurden und die den in seinem Wesen zufälligen Prozeß überlagern, oder durch die Wirkung irgendwelcher Kontroll- und Regelmechanismen, die den Wortzuwachs im Text mit nicht absoluter Genauigkeit steuern, erklärt werden. Die Argumente für die Existenz eines solchen Mechanismus' werden im letzten Teil dieser Arbeit erörtert.

4. Das dargelegte Modell beruht auf der Hypothese der Texthomogenität (die in reellen Texten nicht erfüllt ist). Aber gerade diese Annahme macht das Modell auch bei der Analyse nichthomogener Stichproben nützlich. Vereinigt man mehrere unähnliche Stichproben mit gleichem Z, so wird in der vereinigten Stichprobe der Wert von Z größer (vgl. die Angaben von Ludin über afghanische Texte {169 - 173}; auch für das "Häufigkeitswörterbuch des Tschechischen", in dem alle Texte vereinigt sind, wird Z größer; vgl. {134} mit {58 - 123}, {135 - 142}). Dies ist nicht der Fall, wenn die vereinigten Texte verhältnismäßig homogen sind. Man kann sich jedoch auch die Situation vorstellen (nicht vorhanden in unserem Material), daß Z in der vereinigten Stichprobe kleiner wird als in den sie konstituierenden Stichproben. Eine solche "übermäßige" Homogenität könnte in dem Fall zustandekommen, wenn die konstituierenden Stichproben nur Varianten eines Textes wären. Diese Eigenschaft macht aus Z einen gewissen Indikator der statistischen Homogenität. Aufgrund ähnlicher Überlegungen kann man z.B. schließen, daß die Unterschiede der Z-Werte in Stichproben, die aus sehr heterogenem Material zusammengestellt wurden (vgl. {156 und 159}), den unterschiedlichen Methoden der Stichprobenerhebung und der Auszählung des Lexikons, die die Autoren dieser Stichproben benutzt haben, zuzuschreiben sind.

3.

Man könnte bei diesen Thesen über die formal beschreibende und "instrumentale" Rolle des dargelegten Modells stehenbleiben, wenn nicht durch die angeführten Zahlen einige neue, teilweise ziemlich unerwartete Gesetzmäßigkeiten hindurchschimmerten.

An erster Stelle fällt die außerordentlich große Streuung der Werte von Z auf: von einigen hundert (in Russischen Bylinen) bis zu 11 Millionen in Joyces "Finegans Wake" {50}. Sogar in Werken eines Autors kann diese Größe beträchtlich schwanken: es reicht, das Z in Puškins "bylinenähnlichen" Texten {1,5} und in seinen Gedichten {47,48} zu vergleichen. Obwohl der Wert von Z = 75000, der in seinem Gesamtwerk festgestellt wurde, in großen Zügen die Häufigkeitsstruktur des Gesamtwerks beschreibt (vgl. die obere Graphik auf der Abb. 2), ist er als eine universelle Charakteristik eines beliebigen Textes Puškins praktisch ungeeignet, da der größte Teil seiner Werke ein von 75000 stark abweichendes Z hat.[14] Auch bei L.N. Tolstoj und J. Joyce ist Z variabel. Daher ist Z weder für die Sprache als ganze, noch für den Autor, sondern nur für den einzelnen Text charakteristisch.

Auf der anderen Seite ist das Material in der Tabelle 1a nach wachsendem Z geordnet.[15] Es ist leicht zu sehen, daß diese Anordnung gleichzeitig, zumindest in groben Zügen, einer chronologischen Anordnung entspricht. Wenn man die gereimte Poesie ausschließt und wenn man nicht das Entstehungsdatum eines Textes,

sondern die "Geschichte der Form", in der der Text geschrieben wurde, in Betracht zieht, dann ist die chronologische Anordnung ziemlich exakt. Das Material in der Tabelle 1a beginnt mit den archaischsten Formen des Typs der russischen Bylinen und endet mit den Werken Šolochovs und Joyces.

Wenn Z als ein konventionelles Maß des relativen Vokabularreichtums betrachtet wird, so kann man schließen, daß sich im Laufe großer Zeiträume eine evidente, wachsende Tendenz zur lexikalischen Konzentration der Texte durchsetzt. Die gereimte Poesie hat die Tendenz die gleichzeitige Prosa zu "überholen". So erscheinen Rustavelis "Der Held im Tigerfell" {27} im Bereich der Prosa von Puškin und Tolstoj und Puškins Verse im Bereich der Prosa von Šolochov und Joyce. Ähnliche Beobachtungen sind schon früher gemacht worden (vgl. z.B. die Arbeiten von Kuraškevič), jedoch nur an beschränktem Material; den diesbezüglichen Verallgemeinerungen fehlte nämlich ein Maß des relativen Vokabularreichtums. Es besteht die Hoffnung, daß mit Zunahme empirischer Daten die erwähnten Beobachtungen zu weiterem Fortschritt in dieser Frage führen werden.

Leider muß man bei diesem Plan auf das Material des Häufigkeitswörterbuchs des Tschechischen (HC) {73 - 82} verzichten, da es nicht bekannt ist, was mit Reim und was in verse libre geschrieben wurde.

Es gibt noch einen Aspekt im Verhalten von Z, der uns höchst interessant erscheint. Da diese Größe dieselbe Dimension wie die Textlänge hat, kann man sie miteinander vergleichen. Auffällig ist, daß in der Mehrzahl der Fälle beide Größen von derselben Größenordnung sind. Im allgemeinen erreichen kurze Texte nicht ihr eigenes Z; lange Texte, deren voller Textumfang Z wesentlich überschreitet, werden von Autoren immer in Teile aufgeteilt (Bände, Bücher u.a.), deren Umfang wiederum von der Größenordnung von Z ist. So beträgt der Umfang von "Krieg und Frieden" etwa 20 Z (N = 472000, Z = 24000), aber der Verfasser hat ihn in 17 Teile aufgeteilt, beide Epiloge eingeschlossen. Ein analoges Bild findet man in der "Auferstehung" (3 Teile), in "Podnjataja celina" (2 Bücher), in "Privalovskie milliony" (5 Teile), in den Romanen von Ju. Smolič (6 Teile), G. Tjutjunnik

(2 Bücher), in "Znamenosci" von A. Gončar (3 Bücher) und im Roman "Chleb i sol'" von Stel'mach (3 Teile).

Dieser offensichtliche Zusammenhang zwischen der Textlänge und dem Zipfschen Umfang brachte uns auf den Gedanken, die Korrelation zwischen x = lg N und y = lg Z zu untersuchen. Die Resultate der Analyse unterschiedlicher Gruppierungen des vorhandenen Materials sind in der Tabelle 5 angegeben.

Die Zeilen I-VI in dieser Tabelle beziehen sich auf die Texte, d.h. als N wurde entweder die volle Textlänge (Zeile I),

Tabelle 5.

Gruppierung des Materials				Korrelationskoeffizient		Lineare Regression y = ax+b		Standardabweichung	
		Text- oder Stichprobengruppe	Zahl der Texte oder Stichproben	ρ_{xy}	σ_z	a	b	σ_x	σ_y
Texte	I	$N_o > 10^4$ ohne Daten des HC	27	0,660	0,109	0,833	0,754	0,498	0,630
	II	Idem, mit Berücksichtigung der Teile	28	0,810	0,065	1,255	1,085	0,404	0,622
	III	Künstlerische Prosa des HC (Gruppen A,C,D)	31	0,760	0,076	1,362	-1,310	0,205	0,368
	IV	Vereinigung der Gruppen II und III	59	0,772	0,053	1,160	-0,533	0,337	0,507
	V	Nichtkünstlerische Texte des HC (E,G,H)	14	0,526	0,194	0,378	2,814	0,294	0,211
	VI	$N_o < 10^4$	15	0,653	0,148	0,937	0,940	0,657	1,005
Stichproben	VII	Beliebige Stichproben ohne HC	40	-0,103	0,142	-0,065	5,044	0,663	0,438
	VIII	Beliebige Stichproben aus dem HC	43	0,260	0,142	0,168	3,990	0,617	0,397
	IX	Vereinigung der Stichproben VII und VIII	83	0,031	0,111	0,017	4,641	0,754	0,424

oder die mittlere Länge des Textteils, falls vom Verfasser unterteilt (Zeile II), genommen. Die Zeilen VII-IX beziehen sich auf beliebige Stichproben und als N wurde der tatsächliche Umfang der untersuchten Stichprobe genommen, d.h. entweder der Umfang des Abschnitts oder der Umfang der Vereinigung mehrerer Texte zu einer Stichprobe. Das Material des Häufigkeitswörterbuchs des Tschechischen (HC) (vgl. Jelínek, Bečka, Tešitelová 1961) dient dabei als Kontrollmaterial zu den westlichen Korpora.

Während zwischen der Textlänge und dem Zipfschen Textumfang in allen Fällen offensichtlich eine deutliche Korrelation besteht, gibt es praktisch keine Korrelation zwischen dem Stichprobenumfang und dem zu der Stichprobe gehörenden Z. Am größten ist der Korrelationskoeffizient in den Fällen, wo man die Unterteilung großer Texte vom Verfasser selbst berücksichtigt (Zeile II). Die Texte, die für das "Häufigkeitswörterbuch des Tschechischen" verarbeitet wurden, weisen auch einen hohen Korrelationskoeffizienten auf (Zeile III); der Gesamtkorrelationskoeffizient für alle großen Texte beträgt 0,772 (Zeile IV). Die lineare Regression $y = 1{,}16x - 0{,}533$ liegt in der untersuchten Zone $4 \leq x \leq 5$ relativ nah zur Winkelhalbierenden des Koordinatensystems, was die im Durchschnitt recht große Nähe von Z zum Textumfang bezeugt.[16)]

Bedeutend kleiner ist die Korrelation für nichtkünstlerische Texte des HC (populäre, wissenschaftliche und politische Literatur) und für Texte, die weniger als 10000 Wortverwendungen enthalten (Zeilen V und VI), jedoch übersteigt der Korrelationskoeffizient auch in diesen Fällen den Wert 0,5. Für die nichtkünstlerischen Texte ist der Regressionskoeffizient der linearen Regression wesentlich kleiner als 1, jedoch erlaubt die geringe Anzahl dieser Texte (insgesamt 14) nicht, den Wert von $\underline{a} = 0{,}378$ als endgültig zu betrachten.

Diese Resultate erklären eine früher gemachte Beobachtung, daß große Texte in ihrem vollen Umfang dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgen (vgl. Nadarejšvili, Orlov 1971; Orlov 1969, 1970a,b, 1976). Es entsteht der Eindruck, als ob der Zipfsche Umfang eine obere Grenze für die Textlänge wäre, d.h. der Text kann sich bis zu diesem Umfang erstrecken, aber er kann

ihn nicht wesentlich überschreiten. Wenn es nötig ist, diese Grenze zu überschreiten, dann muß der Text in Teile derselben Größenordnung wie Z unterteilt werden.

Was kann denn einen Autor "aufhalten", wenn er den Zipfschen Umfang erreicht hat? Die Suche nach einer Größe, die beim Zipfschen Umfang einen universellen, für alle Texte gemeinsamen Wert annimmt und gleichzeitig linguistisch sinnvoll ist, führte zu dem Begriff der differentiellen Geschwindigkeit des Vokabularwachstums.

Seien vom Anfang des Textes N Wörter (tokens) geschrieben (oder beobachtet), unter denen es v unterschiedliche Wörter (types) gibt. Bei den nächsten ΔN tokens soll das Vokabular um Δv types anwachsen. Die Größe $\frac{\Delta N}{N}$ bezeichnen wir als den relativen Textzuwachs und $\frac{\Delta v}{v}$ als den relativen Vokabularzuwachs. Das Verhältnis $\frac{\Delta v}{v}/\frac{\Delta N}{N}$ bei kleinem ΔN werden wir die differentielle Geschwindigkeit des Vokabularwachstums nennen (DGVW).

Da wir über die stetige theoretische Vokabularwachstumskurve (10b) verfügen, ist es zweckmäßig, die DGVW durch Differenzierung abzuleiten. Es ist

$$L(X) = \lim \frac{N}{v} \cdot \frac{\Delta v}{\Delta N} = \frac{N}{v} \frac{dv(N,Z)}{dN} = \frac{1}{X-1} \cdot \frac{1}{\ln X} \qquad (17)$$

wo $X = Z/N$.

Es ist interessant, daß diese Größe offensichtlich weder von N, noch von Z, noch von p_{max}, sondern ausschließlich von $X = Z/N$ abhängt. Daraus folgt: Wenn man die Textlänge in Einheiten von der Größe Z ausdrückt, dann erweist sich die Kurve der DGVW als eine universelle, für alle Texte gemeinsame Größe. Wenn sich der Textumfang dem Zipfschen Umfang nähert, d.h. wenn $X \to 1$, dann $L(X) \to 0{,}5$. Der allgemeine Verlauf dieser Funktion ist auf der Abb. 3 dargestellt (vgl. die obere Graphik).

Direkt unter dieser Kurve sind in demselben horizontalen Maßstab die Histogramme der Verteilung der Beziehung $X_o = Z/N_o$ für die untersuchten Texte aufgetragen. Hier ist N_o entweder die volle Textlänge oder die mittlere Textlänge, wenn ein großer Text vom Autor unterteilt wurde. Zwecks Vergleichs wurden darunter die Histogramme der Beziehung $X = Z/N$ für beliebige Stichproben (Ab-

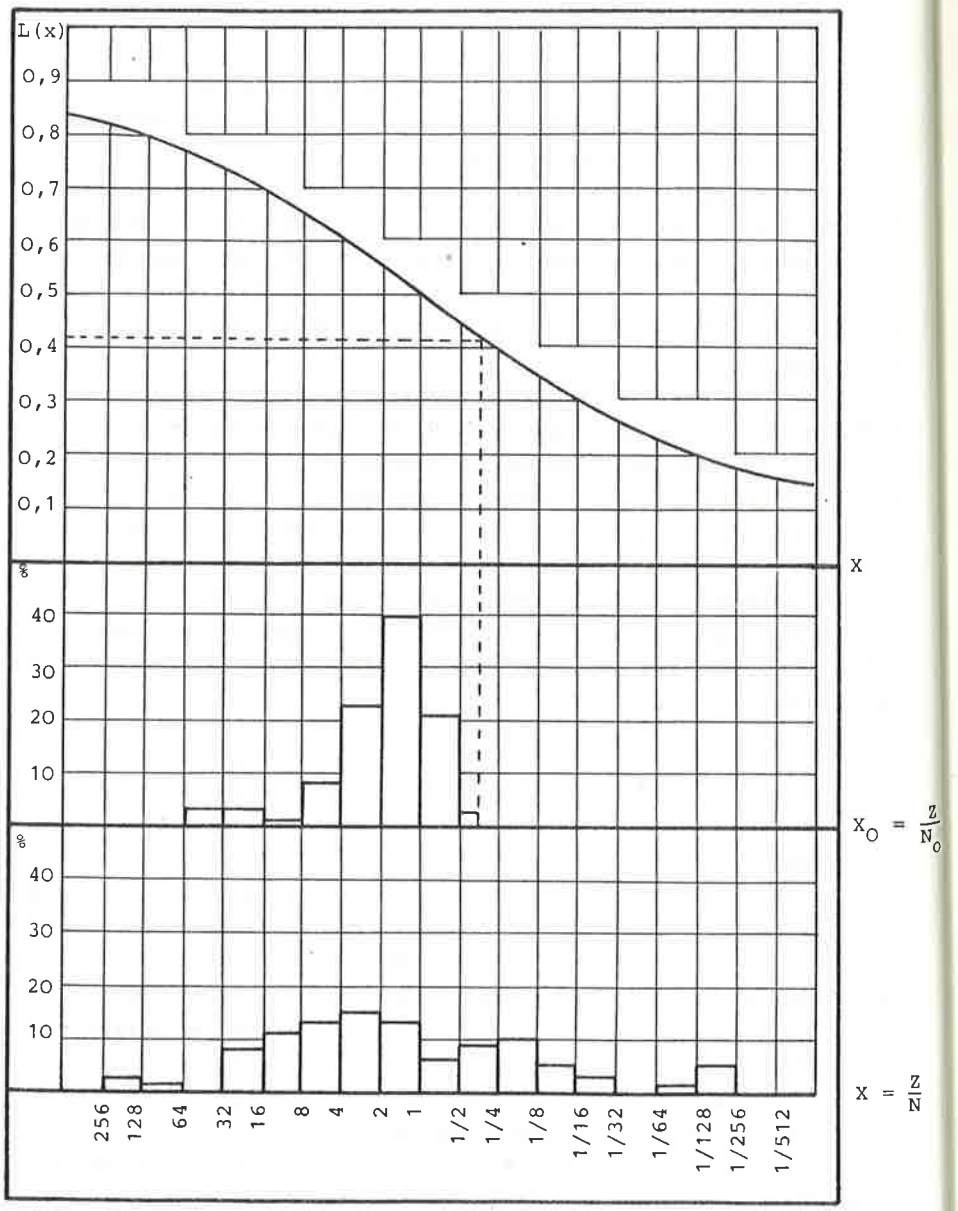


Abb. 3. Graphik der Funktion der DGVW L(X) und die Verteilungen der Größen $X_O = Z/N_O$, wo N_O die volle Textlänge (oder die Länge eines vom Autor bestimmten Teils eines großen Textes) ist, und $X = Z/N$, wo N die Länge einer beliebigen Stichprobe (Abschnitt eines Textes oder Vereinigung mehrerer Texte in einer Stichprobe) ist. Die unterbrochene Linie bezeichnet den kleinsten beobachteten Wert von X_O und $L(X_O)$ für einzelne Texte.

schnitte oder Vereinigungen mehrerer Texte) gezeichnet, und zwar in solchen Fällen, wo der Stichprobenumfang durch Zufall oder durch die Willkür der Forscher festgelegt wurde (z.B. Gesamtwerke von Puškin).

Man kann leicht sehen, während die Verteilung von X_O ein deutliches Maximum in der Nähe von X = 1 (d.h. bei N = Z) hat, besitzt die Verteilung von X kein so deutliches Maximum und ist außerdem beträchtlich auseinandergezogen. 40% aller untersuchten Texte haben eine Länge von 0,5Z bis Z, 62% der Texte von 0,5Z bis 2Z und 84% der Texte von 0,25Z bis 2Z; die Grenzen von 0,25Z bis 2Z entsprechen einer DGVW am Ende des Textes zwischen 0,62 und 0,44. Der niedrigste beobachtete Wert L = 0,414 befindet sich im altgeorgischen Text des "Johannes-Evangeliums". Die DGVW scheint also ein recht begrenzter Parameter zu sein.

Wir werden L(1) = 0,5 als den "nominellen" Wert der DGVW, der die Notwendigkeit der "Textbeendung" signalisiert, betrachten, und wir analysieren die Texte, deren Umfang Z überschreitet. Der mittlere Wert X_O für diese Texte (in unserem Material gibt es 17 dieser Art) ist ungefähr 0,71. Dies entspricht der DGVW L(0,71) = 0,46. Der "nominelle" Wert 0,5 wird also im Schnitt um 8% reduziert, und die größte beobachtete Reduktion ist 17%, wenn eine Reduktion überhaupt stattfindet. Diese Werte liegen in den Grenzen der gewöhnlichen menschlichen Empfindlichkeit für Veränderungen irgendwelcher physikalischer Größen (das Weber-Fechnersche Gesetz).[17]

Daher bietet sich die folgende Hypothese an: bei der Erzeugung oder der Wahrnehmung eines Textes in dieser oder jener Form verwirklicht sich (offenbar unbewußt) eine Kontrolle der DGVW, und die Stelle, wo die Textlänge anfängt mehr als doppelt so schnell zu wachsen wie das Vokabular, wird als Vollendung, Abgeschlossenheit des Textes erlebt.

Ein weiteres Zurückbleiben des Vokabularwachstums hinter dem Textwachstum wird als "Langatmigkeit" des Textes empfunden usw. Außer durch die dargelegten statistischen Erläuterungen kann diese Hypothese auch durch die Beobachtung unterstützt werden, daß bei großen Schriftstellern die volle Textlänge nah an Z liegt, z.B. bei Š. Rustaveli {27}, L.N. Tolstoj {19, 33}, K. Čapek {70},

in der Byline "Vol'ga i Mikula" {3}. Besonders interessant ist die praktisch exakte Übereinstimmung der Textlänge mit dem Zipfschen Umfang in Joyces "Stephen Hero" {53}.

Wie bekannt (vgl. Žantieva 1967) hat dieser frühe Roman von Joyce ursprünglich einen größeren Umfang gehabt, aber der Verfasser hat einen Teil verbrannt. Offensichtlich, wenn der Rest mit Z übereinstimmen sollte, dann müßte der ursprüngliche Umfang Z überstiegen haben. War es vielleicht eben dieser Umstand, der als "Langatmigkeit" empfunden wurde und Joyce beunruhigte? Man kann nicht ausschließen, daß die "Kollisionen mit dem Zipfschen Umfang" für die Autoren eine dramatische Rolle spielen können.

Interessant ist beispielsweise K.G. Paustovskijs Aussage (erwähnt von A. Ionov in "Ogonek" Nr. 22, 1972), daß er nicht imstande ist, ein Werk länger als 11 Druckbogen zu schreiben: "Es ist direkt eine verhexte Zahl: elf, was du auch immer willst! Wenn ich auf den zwölften übergehe, kann ich keine einzige Zeile mehr schreiben". Elf Druckbögen enthalten etwa 50000 Wortverwendungen und in der Tat haben die meisten großen Werke Paustovskijs ungefähr diesen Umfang (das größte Werk "Dalekie gody" hat 14,3 Druckbögen. Nach den Angaben von L. Sudovičene (Učennye zapiski vyss. uč. zav. Lit. SSR. Jazykoznanie 22/2, Vilnjus 1971) hat der Roman "Dym otečestva" den Umfang von N_o = 57000 Wortverwendungen, den Wortschatz von v = 7829 Wörtern und p_1 = 0,0337 (12,35 Druckbögen). Der Wert Z, berechnet aus diesen Angaben, beträgt 65400, d.h. er liegt sehr nah bei der tatsächlichen Romanlänge. Es wäre interessant, die lexikalische Konzentration in anderen Texten von Paustovskij zu untersuchen.

Der parallele Zuwachs des relativen Vokabularreichtums und der maximal "zugelassenen" Umfänge ruft offensichtlich eine Krisensituation hervor. Auf jeden Fall kann man die in der Literaturwissenschaft umstrittene Lebensfähigkeit der Romanform in Begriffen der vorliegenden Arbeit vollständig interpretieren. Der Mensch kann offenbar lexikalisch gesättigte Texte großen Umfangs schwer bewältigen und dieses Gefühl der Erschöpfung der Aufnahmefähigkeit der Leser beunruhigt die Schriftsteller. Anscheinend führte die Entwicklung des Romangenres zur Entstehung

der "literarischen Dinosaurier" ... Künftige Untersuchungen werden zeigen, ob diese Behauptung stimmt.

Zusammenfassend kann man bemerken, daß das dargelegte Modell einige unerwartete Erscheinungen zu beschreiben erlaubt:

(a) das Anwachsen des relativen Vokabularreichtums im Verlauf großer Zeiträume;

(b) das vorauseilende Anwachsen des relativen Vokabularreichtums in der gereimten Poesie;

(c) die Korrelation zwischen der Länge und dem Zipfschen Umfang des Texts;

(d) die Universalität der Funktion der DGVW, wenn die Textlänge in relativierten Einheiten von Z ausgedrückt ist;

(e) die Unzulässigkeit einer wesentlichen Unterschreitung der 0,5-Grenze durch die DGVW (der kleinste beobachtete Wert in zusammenhängendem Text war 0,414).

Offensichtlich hängen (c) und (e) eng zusammen und das eine bedingt das andere. Es ist schwer zu sagen, welche von ihnen primär ist, die Hypothese, daß (e) primär ist, ist wahrscheinlicher. In jedem Fall führt uns der Zusammenhang zwischen N und Z zu literaturwissenschaftlichen Problemen, zu Problemen der Psychologie künstlerischer Kreativität (und zur Psychologie der künstlerischen Rezeption, da man sich schwer vorstellen kann, daß die Autoren diesen Zusammenhang nur zum "eigenen Vergnügen" verwirklichen) und, letzten Endes, zu tieferen Problemen der Organisation und Rezeption von Information durch den Menschen. Es ist offensichtlich, daß dieser Problemkomplex eine weitere vertiefte Untersuchung durch viele Spezialisten, darunter Mathematiker, Linguisten und Psychologen, benötigt.

An dieser Stelle möchte ich meine tiefe Dankbarkeit an A.N. Kolmogorov ausdrücken, dessen Interesse diese Arbeit stark angeregt hatte. Ich bedanke mich herzlich bei V.M. Andrjušenko, N.P. Darčuk, I.Š. Nadarejšvili, T.I. Pataraja und A.Ja. Šajkevič, die mir eine Menge linguistischer Zählungen zur Verfügung stellten.

TABELLE 1A.

	Text	Quelle	Volle Textlänge N_0	Stichprobenumfang N	p_{max}	Wortschatz $\hat{V}$	$R=\frac{V}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
1	Puškin, Iz byta povolžskych razbojnikov /Aus dem Leben der Wolga-Räuber/	Zählung von I.Š. Nadarejšvili	278	278	0,040	112	6,72	400
2	Byline "Brat'ja-razbojniki i sestra" /Räuberbrüder und die Schwester/	- " -	293	293	0,044	116	6,77	486
3	Byline "Vol'ga i Mikula" /Wolga und Mikula/	- " -	930	930	0,040	254	8,32	960
4	Byline "Il'ja Muromec i Kalincar" /Ilja Muromec und Zar Kalin/	- " -	3380	3380	0,059	550	9,46	2130
5	Puškin, Skazka o rybake i rybke /Das Märchen vom Fischer und Fischlein/	Mitteilung von A.Ja. Šajkevič	948	948	0,038	318	10,3	2200
6	Altgeorgisches Evangelium (voller Text)	Imnajšvili (1948)	53043	53043	0,120	2672	11,6	8500
7	Altgeorgisches Johannes Evangelium	- " -	11908	11908	0,084	1146	10,5	4156
8	Altgeorgisches Matthäus Evangelium	- " -	14317	14317	0,115	1708	14,3	10000
9	Altgeorgisches Markus Evangelium	- " -	9528	9528	0,142	1400	14,3	11300

Theoretischer Wortschatz $v(N,Z)$	$x_0=\frac{Z}{N_0}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	$\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
113	1,44	39	60	0,65	36	19	2,00	9	10	0,90
115	1,66	52	61	0,85	27	19	1,40	8	9	0,89
253	1,03	116	127	0,91	56	42	1,33	22	21	1,05
547	0,63	245	258	0,95	82	85	0,97	42	40	1,05
316	2,32	192	178	1,08	45	52	0,87	21	23	0,91
2675	0,16	769	947	0,81	416	400	1,04	242	234	1,03
1147	0,35	432	475	0,91	184	184	1,00	98	99	0,99
1690	0,70	655	789	0,83	277	273	1,01	145	145	1,00
1403	1,22	587	722	0,80	244	234	1,04	133	115	1,16

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{V}$	$R = \frac{V}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
10	Altgeorgisches Lukas Evangelium	Imnajšvili (1948)	17290	17290	0,136	1932	14,7	12000
11	"Čtenie o Borise i Glebe"	Vjalkina & Lukina	7380	7380	0,077	1249	14,5	9000
12	"Žitie Feodosija Pečerskogo"	- " -	18754	18754	0,081	2164	16,4	12000
13	"Skazanie o Borise i Glebe"	- " -	8590	8590	0,107	1611	17,4	18000
14	"Slovo o Polku Igoreve"	Zählung von A.J. Pataraja	2772	2772	0,032	893	16,9	12500
15	Puškin, Skazka o care Saltane /Das Märchen von Zar Salton/	Mitteilung von A.Ja. Šajkevič	~4000	1000	0,041	446	14,1	13500
16	Kumsiašvili, Die Portweinproduktion in Georgien (in Georgisch)	Zählung von I.S. Nadarejšvili	9998	9998	0,051	1906	19,1	17000
17	- " -	- " -	9998	6000	0,056	1484	19,2	20000
18	- " -	- " -	9998	3000	0,068	943	17,2	22400
19	Tolstoj, Krejcerova sonata /Kreutzersonate/	- " -	~25200	10000	0,047	2083	20,8	22000
20	Tolstoj, Vojna i mir /Krieg und Frieden/	- " -	~472000	10000	0,045	2146	21,5	24000
21	Unsuri, Divan (in Arabisch)	Osmanov (1970)	46472	46472	0,048	4824	22,4	26400
22	Puškin, Skazka o zolotom petuške /Das Märchen vom goldenen Hahn/	Mitteilung von A.Ja. Šajkevič	902	902	0,037	457	15,2	26000

Theoretischer Wortschatz $V(N,Z)$	$x_o = \frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	$\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	$\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	$\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
1930	0,69	813	908	0,90	301	321	0,92	177	166	0,92
1247	1,22	607	643	0,94	199	201	0,99	114	102	1,12
2165	0,64	908	992	0,92	296	352	0,84	184	183	1,00
1605	2,09	855	845	1,01	257	310	0,83	123	78	1,42
890	4,50	516	555	0,93	138	137	1,01	66	59	1,12
446	3,37	293	297	0,99	67	62	1,09	31	24	1,31
1903	1,70	817	1035	0,79	295	318	0,93	160	122	1,31
1488	-	706	890	0,79	239	236	1,01	135	138	1,25
948	-	518	623	0,83	142	140	1,01	98	57	1,72
2080	0,87	1169	1173	0,99	349	340	1,03	157	153	1,03
2145	0,051	1181	1230	0,96	393	350	1,12	150	155	0,97
4830	0,57	2268	2200	1,03	653	798	0,82	397	421	0,94
457	28,8	335	337	0,99	51	56	0,91	24	21	1,14

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
23	Saint-Trond, Gesta abbatum trudoniensium I-VII	Tombeur (1945)	26275	26275	0,038	3957	24,4	30000
24	Puškin, Skazka o pope i rabotnike ego Balde /Das Märchen vom Popen und seinem Arbeiter Bald/	Mitteilung von A.Ja. Šajkevič	910	910	0,034	477	15,8	35000
25	Puškin, Pikovaja dama /Pique Dame/	Zählung von I.S. Nadarejšvili	6861	6861	0,038	1928	23,3	35000
26	"Merilo pravednoe"	Vjalkina & Lukina (1964)	31262	31262	0,062	4311	24,4	36000
27	Rustaveli, Der Held im Tigerfell (in Georgisch)	Osmanov (1980)	42120	42120	0,021	5965	29,0	38000
28	Puškin, Kapitanskaja dočka /Die Kapitänstochter/	Josselson (1953)	29345	29345	0,039	4783	27,9	45000
29	- " -	- " -	29345	10000	0,036	2432	34,3	33400
30	- " -	- " -	29345	5000	0,040	1671	23,6	47000
31	- " -	- " -	29345	5000	0,044	1567	22,2	37300
32	Mamin-Sibirjak, Privalovskie milliony /Die Priwalowschen Millionen/	Genkel' (1966)	127927	127927	0,023	10063	28,2	40000
33	Tolstoj, Kazaki /Die Kosaken/	Zählung von I.S. Nadarejšvili	48100	10000	0,056	2582	25,8	53000

Theoretischer Wortschatz $v(N,Z)$	$x_o = \frac{Z}{N_o}$	Einmalige Wörter			Zweimalige Wörter			Dreimalige Wörter		
		beobachtet $\hat{v}_1$	theoretisch $v_1(N,Z)$	$\frac{\hat{v}_1}{v_1(N,Z)}$	beobachtet $\hat{v}_2$	theoretisch $v_2(N,Z)$	$\frac{\hat{v}_2}{v_2(N,Z)}$	beobachtet $\hat{v}_3$	theoretisch $v_3(N,Z)$	$\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
3980	1,14	1775	2040	0,87	705	666	1,06	367	328	1,12
481	38,4	372	336	1,11	42	57	0,74	20	16	1,24
1930	5,08	1146	1218	0,94	335	296	1,12	129	124	1,04
4330	1,15	2147	2220	0,97	706	723	0,98	334	357	0,94
5960	0,90	2995	2940	1,02	937	996	0,94	392	504	0,78
4780	1,54	2384	2575	0,93	847	802	1,06	433	385	1,13
2430	-	1477	1460	1,01	404	382	1,06	167	172	0,97
1662	-	1133	1120	1,01	236	238	0,99	104	96	1,08
1575	-	927	1033	0,90	281	232	1,21	127	95	1,34
10030	0,31	5093	4020	1,27	-	-	-	-	-	-
2570	1,10	1487	1030	0,91	419	391	1,07	190	165	1,15

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
34	Tolstoj, Voskresenie /Die Auferstehung/	Zählung von I.S. Nadarejšvili	~145000	10000	0,057	2587	25,9	53000
35	Richards, Arifmeticeskie operacii na vyčislitelnych masinach /Arithmetische Operationen auf dem Computer/	Borodin & Mater (1967)	60803	29358	0,045	5075	29,7	57000
36	Appolinaire, Calligrammes		9865	9865	0,026	2941	29,8	60000
37	Dovženko, Poema o more (in Ukrainisch) /Gedicht über das Meer/	Mitteilung von N.L. Darčuk	25837	20000	0,037	4187	29,6	57000
38	Smolić, Mir chižinam, vojna dvorcam (in Ukrainisch) /Friede den Hütten, Krieg den Palästen/	- " -	154682	20000	0,033	4906	34,6	94000
39	Tjutjunnik, Vir (in Ukrainisch)	- " -	171860	20000	0,033	4735	33,5	85000
40	Gončar, Mikita Bratus' (in Ukrainisch)	- " -	15078	15078	0,027	3777	30,8	60000
41	Gončar, Znamenosci (in Ukrainisch)	- " -	121055	20000	0,031	4960	35,0	100000
42	Stel'mach, Chleb i sol' (in Ukrainisch) /Brot und Salz/	- " -	194850	20000	0,039	5116	36,1	120000

Theoretischer Wortschatz $v(N,Z)$	$x_o = \frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	Dreimalige Wörter: $\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
2570	0,34	1934	1630	0,94	408	391	1,04	175	165	1,08
5110	0,94	2500	2830	0,88	861	846	1,02	301	377	1,04
2900	6,1	1676	1820	0,92	476	446	1,07	221	179	1,23
4175	2,20	2420	2455	0,99	683	692	0,98	314	306	1,03
4890	0,61	3054	3060	1,00	773	754	1,03	332	324	1,02
4760	0,49	2953	2930	1,01	727	746	0,97	344	318	1,08
3760	4,00	2483	2300	1,08	608	588	1,03	238	253	0,94
4980	0,82	3097	3160	0,98	737	760	0,97	360	362	0,99
5100	0,62	3221	3260	0,99	818	766	1,07	320	320	1,00

TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz V	$R = \frac{V}{\sqrt{N}}$	Zipfscher Umfang Z
1	2	3	4	5	6	7	8	9
43	Gončar, Krov' ljudskaja - ne vodica (in Ukrainisch) /Menschliches Blut ist kein Wasser/	Mitteilung von N.L. Darčuk	72272	20000	0,038	4791	33,9	95000
44	Šolochov, Podnjataja celina Bd. 1 /Neubruch/	Ljatina (1968)	87883	87883	0,032	12778	43,1	130000
45	Šolochov, Podnjataja celina. Bd. 2 /Neubruch/	- " -	106167	106167	0,038	12762	39,2	106167
46	Šolochov, Podnjataja celina. Gesamttext/Neubruch/	- " -	194035	194035	0,035	18508	41,9	127000
47	Puškin, Ruslan i Ljudmila	Mitteilung von A.Ja. Šajkevič	~11000	1000	0,045	579	18,3	178000
48	Puškin, Evgenij Onegin	- " -	~23000	1000	0,046	590	18,6	240000
49	Joyce, Ulysses	Hart (1963)	260430	260430	~0,07	29899	58,7	360000
50.	Joyce, Finnegans Wake	- " -	218077	218077	~0,07	63924	137	$1{,}1 \cdot 10^7$

Theoretischer Wortschatz $v(N,Z)$	$x_o = \frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	$\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	$\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	$\frac{\hat{v}_3}{v_3(N,Z)}$
10	11	12	13	14	15	16	17	18	19	20
4810	1,23	2890	3000	0,96	777	747	1,04	339	316	1,07
12760	1,48	6326	6820	0,93	2212	2040	1,07	1138	947	1,20
12780	1,00	6010	6390	0,94	2390	2130	1,12	1137	1065	1,07
18550	0,65	8048	8610	0,94	3066	3030	1,01	1688	1400	1,20
580	16,2	438	472	0,93	81	53	1,53	23	18	1,28
592	10,4	470	508	0,93	63	52	1,21	22	16	1,37
30100	1,38	16432	16100	1,02	-	-	-	-	-	-
64000	50,5	51922	49200	1,05	-	-	-	-	-	-

	Text	Quelle	volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$
1	2	3	4	5	6	7	8
10a	Beaumarchais,"Figaros Hochzeit"	Musso (1972)	23829	23829	0,060	2367	15,33
15a	Dante,"Das neue Leben"	Alinei (1971)	14392	14392	0,045	2275	18,96
15b	Cervantes, "Don Quijote" (2 Teile, 99 Kapitel)	Gomez (1962)	357255	357255	0,061	7872	18,17
18a	Shakespeare, "König Lear"	Kvaracchelija (1966)	25471	25471	0,039	3391	21,25
19a	Seneca, "De Clementia"	Seneque (1968a)	8218	8218	0,038	1952	21,53
22a	Povest' vremennych let /Die Erzählung von Übergangsjahren/	Tvorogov (1967)	47371	47371	0,081	4684	21,52
23a	Seneca, "De Brevitate vitae"	Seneque (1968b)	6115	6115	0,032	1811	23,16
32a	Pisarev, "Realisty /Die Realisten/	Bulachov(1969)	48354	48354	0,054	6348	28,87
34a	Griboedov, "Gore ot uma" /Verstand schafft Leiden/	Kunickij (1896)	13246	13246	0,032	3343	29,05
37a	Paustovskij, "Dym otečestva /Rauch des Vaterlandes/	Sudavičene (1971)	57000	57000	0,034	7829	32,79
37b	Mickiewicz, "Herr Tadeusz, I-XII	Sambor (1969)	64510	64510	0,032	9250	36,42
37c	-"-, Buch I	-"-	64510	6587	0,037	2257	27,81
37d	-"-, Bücher I-VI	-"-	64510	34280	0,032	6823	36,85
40a	Gladkov, "Povest' o detstve /Erzählung von der Kindheit/	Sinenko (1973)	127917	127917	0,060	12821	35,85

Zipfscher Umfang Z	Theoretischer Wortschatz v(N,Z)	$\chi_o = \frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	Dreimalige Wörter: $\frac{\hat{v}_3}{v_3(N,Z)}$
9	10	11	12	13	14	15	16	17	18	19	20
10340	2356	0,43	1095	1039	1,05	381	413	0,94	214	256	0,84
15230	2267	1,06	1278	1111	1,15	361	370	0,98	157	185	0,84
17300	7905	0,05	2730	2209	1,24	1099	1016	1,08	634	632	1,00
20170	3392	0,79	1933	1630	1,19	483	565	0,85	226	312	0,72
23000	1945	2,80	995	1137	0,88	371	313	1,19	167	140	1,19
28375	4685	0,60	2187	2143	1,02	730	774	0,94	397	407	0,98
32510	1808	5,32	1052	1145	0,92	303	276	1,10	155	116	1,34
52500	6328	1,09	3092	3077	1,00	1058	1012	1,04	546	502	1,09
56146	3342	4,24	2079	2060	1,01	526	521	1,01	210	225	0,93
63800	7845	1,12	3798	3997	0,95	1277	1307	0,98	705	646	1,09
82530	9241	1,28	4360	4810	0,91	1595	1537	1,04	797	855	0,93
84050	2259	-	1460	1564	0,93	331	311	1,06	137	121	1,13
95000	6815	-	3648	3977	0,92	1143	1098	1,04	537	495	1,08
98000	12862	0,77	5636	6146	0,91	2087	2139	0,98	1211	1099	1,10

ZUSATZ ZU TABELLE 1A (FORTSETZUNG)

	Text	Quelle	Volle Textlänge N_o	Stichprobenumfang N	p_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$
1	2	3	4	5	6	7	8
43a	Dante, "Die göttliche Komödie"	Alinei (1971)	101554	101554	0,039	13004	40,81
43b	Dante, "Die göttliche Komödie, Hölle"	- " -	101554	34126	0,041	6579	35,61
43c	Dante, "Die göttliche Komödie, Fegefeuer"	- " -	101554	34042	0,039	6450	34,96
43d	Dante, "Die göttliche Komödie, Paradies"	- " -	101554	33386	0,039	6273	34,33
46a	Byron, "Don Juan"	Byron (1967)	130745	130745	0,046	14411	39,84

Anmerkung: Der Zusatz zur Tabelle 1a wurde nach der Fertigstellung der deutschen Übersetzung erstellt und konnte nicht in die Tabelle 1 eingeordnet werden. Die Zahlen in der ersten Spalte zeigen, an welche Stelle in der Tabelle 1a die Daten einzuordnen sind.

Zipfscher Umfang Z	Theoretischer Wortschatz $v(N,Z)$	$\chi_o = \frac{Z}{N_o}$	Einmalige Wörter: beobachtet $\hat{v}_1$	Einmalige Wörter: theoretisch $v_1(N,Z)$	Einmalige Wörter: $\frac{\hat{v}_1}{v_1(N,Z)}$	Zweimalige Wörter: beobachtet $\hat{v}_2$	Zweimalige Wörter: theoretisch $v_2(N,Z)$	Zweimalige Wörter: $\frac{\hat{v}_2}{v_2(N,Z)}$	Dreimalige Wörter: beobachtet $\hat{v}_3$	Dreimalige Wörter: theoretisch $v_3(N,Z)$	Dreimalige Wörter: $\frac{\hat{v}_3}{v_3(N,Z)}$
9	10	11	12	13	14	15	16	17	18	19	20
118000	12966	1,16	7487	6645	1,13	1966	2159	0,91	891	1063	0,84
94000	6582	-	4100	3838	1,07	995	1061	0,94	439	478	0,92
86850	6447	-	4009	3720	1,08	937	1044	0,90	399	474	0,84
82200	6269	-	3939	3599	1,09	899	1017	0,88	407	464	0,88
120200	14434	0,92	7250	7512	0,97	2405	2504	0,96	1194	1252	0,95

TABELLE 1B

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
51	Byline "Avdot'ja-ženka rjazanočka"	Zählung von I.S. Nadarejšvili	524	524	0,034	163	7,13	415	0,79
52	La Rochefoucaud, Maximes	Mitteilung von A.Ja. Šajkevič	13500	13500	0,066	1893	16,3	12000	0,96
53	Joyce, Stephen Hero (???)	Hart (1963)	74459	74459	~0,07	8740	31,9	~74500	~1
54	Richards, Areifmeticeskie operacii na vyčislitel'nych mašinach /Arithmetische Operationen auf dem Computer/	Borodin & Mater (1967)	60803	60803	0,045	8978	36,4	~90000	~1,48
55	Macaulay, "Essay on Bacon"	Frumkina (1964)	?	2046	~0,07	964	21,4	~150000	-
56	- " -	- " -	?	4049	~0,07	1432	22,5	~ 65000	-
57	- " -	- " -	?	6045	~0,07	2048	26,4	~100000	-
58	V. Vančura, "Konec starých časů" 7-186	HC, Gruppe A künstlerische Prosa	?	30281	~0,047	5469	31,4	~ 60000	-

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
59	K. Horký, "Piskání v lese" (2)	HC, Gruppe A, künstlerische Prosa	28028	28028	~0,047	6111	36,5	~ 90000	~ 3,22
60	J. Fučík, "Reportáz psana na oprátce" (3)	" - "	21640	21640	-"-	5006	34,0	~ 90000	~ 4,15
61	K. Čapek, "Zivot a dílo skladatele Foltýna (4)	" - "	21963	21963	-"-	4145	27,9	~ 45000	~ 2,06
62	J. Morávek, "Srdce na zámek" 5-171 (5)	" - "	?	35187	-"-	6927	36,9	~100000	-
63	I. Olbracht, "Bratr Žak" (6)	" - "	29803	29803	-"-	5559	32,2	~ 67000	~ 2,25
64	K. Nový, Rytíři a lapkově" 5-185 (7)	" - "	?	33774	-"-	6265	34,1	~ 80000	-
65	J. Spáčil, "Náš svět zemře s námi" (8)	" - "	23436	23436	-"-	4582	29,9	~ 57000	~ 2,43
66	J. Marek, "Vesnice pod zemí" 9-160 (9)	" - "	?	31195	-"-	4188	23,7	~ 48000	-
67	E. Bass, "Lidé z maringotek" (11)	" - "	47542	47542	-"-	8763	39,8	~120000	~ 2,53

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
68	J. Hora, "Hladový rok" (12)	HC, Gruppe A, künstlerische Prosa	35273	35273	~0,047	6939	36,9	~100000	~ 2,88
69	M. Pujmanová, "Předtucha" (13)	- " -	24353	24353	-"-	4858	31,1	~ 63000	~ 2,59
70	K. Čapek, "Obyčejný život" (14)	- " -	39360	39360	-"-	5539	27,9	~ 40000	~ 1,02
71	J. Marânek, "Barbar Vok" (15)	- " -	30145	30145	-"-	6498	37,4	~120000	~ 3,98
72	V. Řezáč, "Černě svetlo" (16)	- " -	55164	55164	-"-	8675	36,9	~ 95000	~ 1,73
73	Vl. Holan, "Havranim brkem"(31)	HC, Gruppe B, Poesie	?	12153	-"-	3855	35,0	~170000	-
74	J. Seifert, "Jaro sbohem" (32)	- " -	?	6171	-"-	2095	25,5	~ 55000	-
75	S.K. Neumann, "Zamořená leta" (33)	- " -	?	6658	-"-	2435	29,8	~130000	-
76	J. Hora, "Kniha domova" (34)	- " -	?	10900	-"-	2961	28,4	~ 60000	-
77	Er. Halas, "Ladeni" (35)	- " -	?	5249	-"-	2078	28,7	~150000	-
78	J. Hořec, "Květen Č. I" (36)	- " -	?	3374	-"-	1204	20,7	~ 35000	-

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
79	V. Nezval, "Historický obraz" I-III (37)	HC, Gruppe B, Poesie	?	5340	~0,047	1825	25,0	~ 60000	-
80	P. Křička, "Svetlý oblak" (38)	- " -	?	4782	-"-	1689	24,4	~ 60000	-
81	J. Hirśal, "Student nebe" (39)	- " -	?	3614	-"-	1066	17,7	~ 18000	-
82	Fr. Hrubín, "Krásno po chudobě" (40)	- " -	?	2240	-"-	914	19,3	~ 36000	-
83	M. Majerová, "Robinsonka" (41)	HC, Gruppe C, Jugendliteratur	37509	37509	~0,043	7420	38,3	~120000	~ 3,20
84	J. Prchal, "Bílý jestráb" (42)	- " -	13268	-"-	-"-	2761	24,0	~ 30000	~ 2.44
85	J. Pilař, "Sluneční vuz" (43)		18400	-"-	-"-	4164	30,7	~ 60000	~ 3,26
86	J. John, "Narodil se.." (44)	- " -	24779	-"-	-"-	6286	39,3	~150000	~ 6,07
87	Fr. Langer, "Deti a dýka" (45)	- " -	34192	34192	-"-	5599	30,3	~ 55000	~ 1,61
88	J.V. Pleva, "Budík" (46)	- " -	24249	24249	-"-	3120	20,0	~ 17000	~ 0,70

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
89	B. Říha, "Na útěku" (47)	HC, Gruppe C, Jugendliteratur	30072	30072	~0,043	4562	26,4	~ 37000	~ 1,24
90	E. Štorch, "Meč proti meči" (48)	- " -	52031	52031	-"-	7350	32,2	~ 70000	~ 1,34
91	V. Deyl, "Vyzvědačí"(49)	- " -	31079	31079	-"-	5680	32,2	~ 67000	~ 2,16
92	J. Kopta, "Přivor pod ořechy" (50)	- " -	32284	32384	-"-	6011	33,4	~ 75000	~ 2,39
93	M. Kratochvil, "České jaro" (51)	HC, Gruppe D,	12384	12384	~0,031	2156	19,4	~ 15000	~ 1,21
94	Er. Tetauer, "Krivdu napravovatí" (52)	- " -	16250	16250	-"-	2324	18,1	~ 13000	~ 0,79
95	J. a M. Tomanovi,"Vinice" (53)	- " -	12444	12444	-"-	1999	17,9	~ 12000	~ 1,09
96	O. Scheinpflugová, "Guayana" (54)	- " -	14694	14694	-"-	2417	19,9	~ 16000	~ 1,09
97	Er. Gotz, "Soupeři" (55)	- " -	15400	15400	-"-	2817	22,6	~ 24000	~ 1,56
98	St. Lom, "Karel IV" (56)	- " -	18683	18683	-"-	3840	28,1	~ 50000	~ 2,68
99	L. Suchý, "Hrstka věrných" (57)	- " -	8910	8910	-"-	1790	18,9	~ 15000	~ 1,68

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
100	J. Klíma, "Na dosah ruky" (58)	HC, Gruppe D,	12852	12852	~0,031	1480	13,0	~ 5000	~ 0,39
101	J. Drda, "Hrátky s čertem" (59)	- " -	13908	13908	-"-	2899	24,6	~ 30000	~ 2,16
102	Fr. Langer, "Jiskra v popelu" (60)	- " -	14418	14418	-"-	2661	22,1	~ 22000	~ 1,52
103	V. Přihoda, "Idem na školu II stupne" (61)	HC, Gruppe E,	24658	24658	~0,038	4308	27,5	~ 47000	~ 1,91
104	B. Václavek, "Lidová slovesnost" (62)	- " -	9290	9290	-"-	2360	24,5	~ 36000	~ 3,88
105	A. Piša, "Poesie nově doby" (63)	- " -	23802	13802	-"-	3381	21,9	~ 22000	~ 0,92
106	"Ústava Československé republiky" (65)	- " -	8374	8374	-"-	1337	14,6	~ 6000	~ 0,72
107	K. Chochola, "Spalovací motory" (66)	- " -	31655	31655	-"-	3388	19,0	~ 15000	~ 0,47
108	V. Vojtíšek, "Česká města" (67)	- " -	8729	8729	-"-	2116	22,6	~ 28000	~ 3,21
109	B. Gloss, "Metoda plánování práce" (68)	- " -	18085	18085	-"-	2372	17,6	~ 13000	~ 0,72

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
110	P. Reiman, "Století vědeckého socialismu" (69)	HC, Gruppe E,	10103	10103	~0,038	2831	28,1	~ 60000	~ 5,94
111	Zd. Nejedlý, "Dějiny národa českého" (81)	HC, Gruppe C, Jugendliteratur	31250	31250	~0,035	4366	24,7	~ 30000	~ 0,96
112	J. Ulrich, "Zaklady marxistické ekonomie" (82)	- " -	?	26908	-"-	3577	21,8	~ 22000	-
113	V. Úlehla, "Zamyšleni nad životem II" (83)	- " -	?	20340	-"-	3870	27,1	~ 40000	-
114	O. Chlup, Pedagogika(84)	- " -	?	29813	-"-	4516	26,1	~ 35000	-
115	J. Janko, "Základy statistické indukce" (85)	- " -	?	17249	-"-	1835	13,9	~ 6000	-
116	K. Přerovský. Therapie uhlečitá (86)	- " -	18?	18448	-"-	3088	22,7	~ 24000	-
117	R. Souček, "Psychoanalisa," (87)	- " -	20603	20603	-"-	3502	24,4	~ 28000	1,36
118	K. Honzík, "Tvorba životního slohu" (88)	- " -	?	33700	-"-	5916	32,2	~ 65000	-
119	V. Láska, "Úvod do geofysiky" (89)	- " -	14714	14714	-"-	2790	23,0	~ 26000	1,76

TABELLE 1B (FORTSETZUNG)

	Text	Quelle	N_o	N	P_{max}	$\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Z	$X_o = \frac{Z}{N_o}$
	2	3	4	5	6	7	8	9	10
120	J. Popelová, "Tri studie z filosofie dejin" (90)	HC, Gruppe C, Jugendliteratur	?	23237	~0,035	3970	26,0	~ 25000	-
121	Zd. Nejedlý, "O kulturu národní a lidovou" (92)	HC, Gruppe H,	17579	17579	~0,050	3137	23,6	~ 27000	~ 1,53
122	E. Burian, "Voláno rozhlasem" (93)	- " -	22420	22420	-"-	3886	26,0	~ 36000	~ 1,60
123	A. Zápotocký, "Po staru se žít nedá" (94)	- " -	30797	30797	-"-	4849	27,6	~ 40000	~ 1,30

	Text	Quelle	Stichprobenumfang N	P_{max}	Wortschatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipfscher Umfang Z
124	Častotnyj slovar' voennych tekstov /Häufigkeitswörterbuch der Texte aus dem Militärwesen/	Kolguškin (1970)	689214	0,029	3000	3,62	2150
125	Russische Texte über Elektronik	Kalinina (1968)	200388	0,033	6826	15,3	15200
126	Častotnyj slovar' jazyka Abaja /Häufigkeitswörterbuch der Sprache von Abaj/	Bektaev (1966)	46847	~0,05?	6017	27,8	46847
127	Puškin, Gesamtwerk	Materialy (1963)	544777	0,046	21197	30,0	75000
128	Abschnitte aus dem Werk von V. Pšavela (in Georgisch)	Abuladze (1966)	26467	0,046	5656	34,8	100000
129	- " -	- " -	18187	0,044	4593	34,1	105000
130	- " -	- " -	8280	0,050	2681	29,5	100000
131	Häufigkeitswörterbuch des Russischen	Zasorina (1966)	120474	0,044	14208	40,8	130000
132	Lenins Werke, Bd.1, 1-130	Borodin & Mater (1967)	27110	0,039	6250	38,0	130000
133	Häufigkeitswörterbuch der modernen ukrainischen Prosa	(1969)	100000	~0,035	13945	44,2	142000
134	Häufigkeitswörterbuch des Čechischen	Jelínek & Bečka & Tešitelová (1961)	1623527	0,0413	54486	42,7	210000

$x = \frac{Z}{N}$	Theorethischer Wortschatz v(N,Z)	Einmalige Wörter $\hat{v}_1$	v(1,N,Z)	$\frac{\hat{v}_1}{v(1,N,Z)}$	Zweimalige Wörter $\hat{v}_2$	v(2,N,Z)	$\frac{\hat{v}_2}{v(2,N,Z)}$	Dreimalige Wörter $\hat{v}_3$	v(3,N,Z)	$\frac{\hat{v}_3}{v(3,N,Z)}$
0,0031	3010	712	512	0,72	284	253	1,12	150	166	0,90
0,076	6800	2009	2080	0,96	863	932	0,93	536	570	0,94
~1	6050	2975	3025	0,98	902	1010	0,89	491	504	0,97
0,138	21200	6388	7300	0,88	2910	3140	0,93	1803	1845	0,96
3,38	5670	-	3450	-	-	890	-	-	390	-
5,77	4570	2714	2920	0,93	733	675	0,92	321	291	1,10
12,1	2665	1675	1840	0,91	418	370	1,13	154	145	1,06
1,08	14200	6752	7200	0,96	2337	2380	0,98	1171	1180	0,99
4,8	6270	3607	3920	0,92	1013	985	1,03	500	410	0,82
1,42	13950	6679	7350	0,91	2220	2300	0,97	1147	1135	1,01
0,129	54500	20476	18530	1,10	7762	7950	0,98	-	-	-

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
135	Häufigkeitswörterbuch des Čechischen, biologische Texte	Jelínek & Bečka & Tešitelová (1961)	32972	~0,038	4750	26,2	~ 36000	1,09
136	-"-, Zeitschrift Tvorba 27,35	- " -	26725	~0,040	5477	33,4	~ 75000	2,78
137	-"-, Zeitung Práce 30.3.1950	- " -	24903	-"-	5315	33,6	~ 80000	3,21
138	-"-, Zeitung Právo lidu 4.9.1946	- " -	13049	-"-	3718	32,6	~ 90000	6,67
139	-"-, Zeitschrift Svět práce 5,36	- " -	32079	-"-	6782	37,8	~120000	3,74
140	-"-, Zeitung Svobodné slovo 20.9.1946	- " -	12574	-"-	3742	33,4	~120000	9,55
141	-"-, Zeitung Rudé právo 5.9.1946	- " -	14502	-"-	4166	34,6	~130000	8,95
142	-"-, Lidová demokracie 11.9.1946	- " -	13369	-"-	4160	36,1	~140000	10,47
143	Čechisch (nach Vey)	Novak (1962)	1523627	~0,05	55795	45,1	~250000	0,164
144	Russische Umgangs-sprache	Ovsienko (1966)	100000	~0,05	6300	19,9	~ 25000	0,25

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
145	Häufigkeitswörterbuch des Russischen	Štejnfel'd (1963)	400000	0,046	24224	38,3	122000	0,30
146	Russische wissenschaftliche Texte	Ludin (1971)	50000	~0,05	9464	42,3	~170000	3,40
147	Russische Umgangssprache	- " -	50000	~0,05	10304	46,2	~230000	4,60
148	Französisch	Novak (1962)	400000	~0,07	7628	12,1	~ 17000	0,042
149	Französische Umgangs-sprache	Novak (1962)	312135	~0,07	7995	14,3	~ 20000	0,064
150	Französische Umgangs-sprache	Ovsienko (1966)	312200	~0,07	8000	14,3	~ 20000	0,064
151	Rumänisch, Belletristik	Novak (1962)	50000	~0,035	4547	20,3	~ 20000	0,40
152	Rumänisch, Publizistik	- " -	70000	~0,035	5710	21,6	~ 25000	0,36
153	Häufigkeitswörterbuch des Rumänischen und des Moldauischen	- " -	300416	0,035	14250	25,8	~ 55000	0,183
154	Englische Telephongespräche (French)	Alekseev (1971)	79300	~0,07	2240	7,95	~ 4250	0,058
155	- " -	- " -	80000	~0,07	2400	8,50	~ 5000	0,063

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
156	Englische Umgangssprache	Alekseev (1971)	512647	~0,07	4539	6,33	~ 6300	0,012
157	Englische Korrespondenz	- " -	23629	~0,07	2001	13,0	~ 8000	0,340
158	Englische Korrespondenz (Cook, O'Shea)	- " -	200000	~0,07	5200	11,6	~ 12000	0,060
159	Englische Umgangssprache	- " -	288000	~0,07	6800	12,5	~ 15000	0,052
160	Englische Elektronik	Alekseev (1969)	200000	0,10	7160	16,0	21000	0,105
161	Englische Texte über Automobilbau	Budman (1971)	100000	0,105	5200	16,4	20000	0,20
162	- " -	- " -	200000	0,107	7355	16,5	24000	0,12
163	Engl. wissensch. Texte	Ludin (1971)	50000	~0,07	5399	24,1	~ 38000	0,76
164	Geschriebenes Englisch der Schüler	Alekseev (1971)	6012359	~0,07	25632	10,45	~ 40000	0,007
165	Englische Zeitungssprache	Ludin (1971	44000	~0,07	6002	28,7	~ 60000	1,37
166	Englische Literatursprache (Dewey)	Alekseev (1971)	100000	~0,07	10161	32,1	~ 80000	0,80
167	Häufigkeitswörterbuch medizinischer Lehrbücher	Levitskij (1966)	82155	0,035	6698	23,4	28300	0,34

TABELLE 2A (FORTSETZUNG)

	Text	Quelle	Stich-proben Umfang N	p_{max}	Wort-schatz $\hat{v}$	$R = \frac{v}{\sqrt{N}}$	Zipf-scher Umfang Z	$X = \frac{Z}{N}$
168	Wörterbuch der organischen Chemie (Zarechnak)	Zasorina (1966)	100000	~0,05	8000	25,3	~ 40000	0,40
169	Afghanisch, Umgangssprache	Ludin (1971)	40000	~0,05?	6411	32,0	~ 70000	1,75
170	-"-, wissenschaftliche Texte	- " -	40000	-"-	6536	32,6	~ 75000	1,87
171	-"-, Zeitungstexte	- " -	40000	-"-	6755	33,8	~ 80000	2,00
172	-"-, künstlerische Prosa	- " -	40000	-"-	7000	35,0	~ 95000	2,38
173	-"-, Poesie	- " -	40000	-"-	7280	36,3	~120000	3,00
174	Gemischte Stichprobe aller afghanischen Texte	- " -	200000	-"-	21268	47,7	~190000	0,95

TABELLE 3

	Text	Nr. der Stichprobe 1	N_1	v_1	Z_1	Nr. der Stichprobe 2 $v(N_1,Z_i)$	Z_2 $(1-\frac{v_1}{v(N_1,Z_2)})100\%$	Theoretisches Vokabular	Relativer Fehler	Zulässige theoretische Abweichung 3σ
1	Puškin, Kapitanskaja dočka /Kapitänstochter/	28	29345	4783	45000	29	33400	4420	+8,21*	± 6,38
2	- " -	28	-"-	-"-	-"-	30	47000	4860	-1,59	±11,73
3	- " -	28	-"-	-"-	-"-	31	37300	4510	+6,05	± 9,81
4	- " -	29	10000	2432	33400	28	45000	2565	-5,19*	± 4,92
5	- " -	29	-"-	-"-	-"-	30	47000	2606	-6,68	± 8,07
6	- " -	29	-"-	-"-	-"-	31	37300	2440	-0,33	± 8,35
7	- " -	30	5000	1671	47000	28	45000	1640	+1,89	± 5,82
8	- " -	30	-"-	-"-	-"-	29	33400	1570	+6,43*	± 6,40
9	- " -	30	-"-	-"-	-"-	31	37300	1575	+6,10	± 7,94
10	- " -	31	5000	1567	37300	28	45000	1640	-4,45	± 5,80
11	- " -	31	-"-	-"-	-"-	29	33400	1570	-0,19	± 5,92
12	- " -	31	-"-	-"-	-"-	30	47000	1662	-5,72	± 9,80
13	Šolochov, Podnjataja celina/Neubruch/	44	87883	12778	130000	45	106167	11600	+10,15*	± 2,37

TABELLE 3 (FORTSETZUNG)

	Text	Nr. der Stichprobe 1	N_1	v_1	Z_1	Nr. der Stichprobe 2 $v(N_1,Z_i)$	Z_2 $(1-\frac{v_1}{v(N_1,Z_2)})100\%$	Theoretisches Vokabular	Relativer Fehler	Zulässige theoretische Abweichung 3σ
14	Šolochov, Podnjataja celina /Neubruch/	44	87883	12778	130000	46	127000	12550	+ 1,82	± 2,06
15	- " -	45	106167	12762	106167	44	130000	14130	- 9,68*	± 2,40
16	- " -	45	-"-	-"-	-"-	46	127000	13800	- 7,52*	± 1,92
17	- " -	46	194035	18508	127000	44	130000	18800	- 1,55	± 2,41
18	- " -	46	-"-	-"-	-"-	45	106167	17000	- 8,87*	± 2,30
19	Englischer Automobilbau	162	100000	5200	20000	163	24000	5750	- 9,57*	± 2,60
20	- " -	163	200000	7355	24000	162	20000	6670	+10,27*	± 3,24
21	Važa Pšavela, Auswahl von Texten	129	26467	5656	100000	129	105000	5860	- 3,52	± 4,40
22	- " -	128	-"-	-"-	-"-	130	100000	5670	- 0,25	± 6,86
23	- " -	128	18187	4593	105000	128	100000	4500	+ 2,06	± 4,13
24	- " -	129	-"-	-"-	-"-	130	100000	4500	+ 2,06	± 6,46
25	- " -	130	8280	2681	100000	128	100000	2665	+ 0,60	± 4,67

TABELLE 3 (FORTSETZUNG)

	Text	Nr. der Stichprobe 1	N_1	v_1	Z_1	Nr. der Stichprobe 2 $v(N_1,Z_1)$	Z_2 $(1-\frac{v_1}{v(N_1,Z_2)})100\%$	Theoretisches vokabular	Relativer Fehler	Zulässige theoretische Abweichung 3σ
26	Važa Pšavela, Auswahl von Texten	130	8280	2681	100000	129	105000	2740	- 2,16	± 5,05
27	Kumsiašvili, Portweinproduktion...	16	9998	1906	17000	17	20000	1990	- 4,22	± 7,75
28	- " -	16	-"-	-"-	-"-	18	22400	1980	- 3,74	± 11,71
29	- " -	17	6000	1484	20000	16	17000	1427	+ 3,99	± 6,66
30	- " -	17	-"-	-"-	-"-	18	22400	1470	+ 0,95	± 10,83
31	- " -	18	3000	943	22400	16	17000	933	+ 1,07	± 7,86
32	- " -	18	-"-	-"-	-"-	17	20000	962	- 1,97	± 8,31
33	Macaulay, "Essay on Bacon"	55	2046	964	150000	56	65000	870	+10,80*	± 9,45
34	- " -	55	-"-	-"-	-"-	57	100000	926	+ 4,11	± 8,79
35	- " -	56	4049	1432	65000	55	150000	1625	-11,88	± 11,90
36	- " -	56	-"-	-"-	-"-	57	100000	1540	- 7,01	± 7,56
37	- " -	57	6045	2048	100000	55	150000	2185	- 6,27	± 12,40
38	- " -	57	-"-	-"-	-"-	56	65000	1880	+ 8,93*	± 8,90

TABELLE 4A

	Pikovaja Dama Z = 35000				Il'ja Muromec i Kalin-car Z = 2130				Slovo o polku Igoreve Z = 12500			
N	v	v(N,Z)	$(1-\frac{v}{v(N,Z)})100\%$	3σ	v	v(N,Z)	$(1-\frac{v}{v(N,Z)})100\%$	3σ	v	v(N,Z)	$(1-\frac{v}{v(N,Z)})100\%$	3σ
250	165	169	-2,37	-	107	124	-13,70	±20,42	159	166	-4,22	±19,40
500	281	293	-4,10	±14,70	175	194	- 9,79	-	277	279	-0,72	-
1000	462	497	-7,05	-	257	291	-11,69	±16,75	461	457	+0,88	-
1500	619	685	-9,63*	± 9,60	369	364	+ 1,37	-	613	602	+1,83	±10,78
2000	787	845	-6,87	± 8,13	419	423	- 0,94	-	737	725	+1,65	-
2500	934	985	-5,08	-	461	444	- 3,83	±11,04	842	840	+0,24	±10,35
3000	1067	1120	-4,73	± 7,77	512	518	- 1,16					
4000	1348	1360	-0,87									
5000	1541	1575	-2,16									
6000	1752	1770	-1,02									

TABELLE 4B

N	Krejcerova Sonata Z = 22000			Kazaki Z = 53000			Vojna i mir Z = 24000			Voskresenie Z = 53000		
	v	$\wedge(Z'N)$	$(1-\frac{\wedge(Z'N)}{v})100\%$	v	$\wedge(Z'N)$	$(1-\frac{\wedge(Z'N)}{v})100\%$	v	$\wedge(Z'N)$	$(1-\frac{\wedge(Z'N)}{v})100\%$	v	$\wedge(Z'N)$	$(1-\frac{\wedge(Z'N)}{v})100\%$
1000	431	445	- 3,15	438	506	-13,42*	454	474	-4,22	489	506	-3,36
2000	685	760	- 9,87*	732	850	-13,90*	759	774	-1,94	824	850	-3,06
3000	886	997	-11,12*	1002	1140	-12,12*	1025	1020	+0,49	1132	1140	-0,70
4000	1092	1200	- 9,00*	1233	1395	-11,61*	1218	1230	-0,98	1409	1395	+1,00
5000	1288	1380	- 6,66	1540	1630	- 5,51	1883	1415	-2,26	1640	1630	+0,61
6000	1505	1542	- 2,39	1814	1840	- 1,41	1553	1585	-2,02	1863	1840	+1,25
7000	1682	1695	- 0,77	2048	2040	+ 0,34	1691	1740	-2,82	2049	2040	+0,44
8000	1819	1830	- 0,60	2253	2225	+ 1,26	1830	1881	-2,70	2237	2225	+0,54
9000	1958	1960	- 0,10	2438	2400	+ 1,58	2005	2020	-0,74	2411	2400	+0,46
10000	2048	2080	-	2582	-	-	2146	2145	-	2587	2570	-
			36			36			36			36
			-			±11,15			±11,13			±11,13
			±8,64			± 8,53			-			-
			±7,64			± 7,50			-			-
			±7,14			± 6,94			-			-
			±6,77			± 6,55			± 6,70			± 6,70
			-			-			-			-
			-			-			-			-
			-			-			± 6,23			± 6,23
			-			-			-			-
			-			-			-			-

ANHANG I

EINE APPROXIMATIVE METHODE ZUR BERECHNUNG DES ZIPFSCHEN UMFANGS

Gleichungen wie (13) löst man üblicherweise durch iterative Approximation. Damit das Verfahren beginnen kann, muß man die erste Approximation, den groben Orientierungswert der Gleichungswurzel kennen.

Eine gute erste Approximation kann man erhalten, wenn man Gleichung (13) graphisch löst. Die beigefügten Nomogramme erlauben, eine Übereinstimmung der linken und rechten Seite von (13) mit höchstens 3-5% Abweichung zu erreichen.

Jede der Kurven auf Abb. 4-8 stellt den Quotienten von Guiraud für einen homogenen Text mit festgelegten Werten Z und p_{max} dar:

$$R(N) = \frac{v(N,Z)}{\sqrt{N}}.$$

Es ist leicht zu sehen, daß diese Beziehung am Anfang des Textes wächst und dann anfängt abzunehmen. Gerade wegen dieses Umstandes ist diese Beziehung als Maß des relativen Vokabularreichtums ungeeignet.

Die stark schiefe Funktion

$$R_2(N) = \frac{v(N)}{\sqrt{N}}$$

stellt eine analoge Beziehung für Texte dar, die bei ihrem Umfang dem verallgemeinerten Zipf-Mandelbrotschen Gesetz folgen, mit anderen Worten, jeder Punkt dieser Kurve entspricht einem Text mit N = Z, und der Zähler dieser Beziehung wird laut (7) berechnet. Daraus ist evident, daß die Abzissen der Schnittpunkte von R(N) und $R_Z(N)$ gleich den Werten von Z für die Kurve R(N) sind. (Aus diesem Grund wurden die Werte von Z, die jede Kurve charakterisieren, auf die Graphiken nicht aufgetragen.) Es ist interessant, daß die Kurve $R_2(N)$ durch die Maxima der Kurven

R(N) läuft: das bedeutet, daß Guirauds Beziehung gerade beim Zipfschen Umfang den für den gegebenen Text höchsten erreichbaren Wert annimmt.

Sei gegeben eine lexikalische Stichprobe mit bekannten N, $\hat{v}$ und p_{max}. Für die Bestimmung von Z berechnen wir $\hat{R} = \frac{\hat{v}}{\sqrt{N}}$ und tragen es als Punkt auf das Nomogramm auf, bei dem der Parameter p_{max} der Häufigkeit des häufigsten Wortes am nächsten steht. Es kann eines von drei Ereignissen zustande kommen:

1. Der Punkt $(N,\hat{R})$, unserer Stichprobe liegt auf der stark schiefen Kurve $R_2(N)$. In dem Falle nehmen wir an, daß $Z = N$ ist.
2. Der Punkt $(N,\hat{R})$ liegt auf einer der konkaven Kurven R(Z). In diesem Falle ist Z als die Abzisse des Schnittpunkts dieser Kurve mit der Kurve $R_2(N)$ gegeben.
3. Am wahrscheinlichsten ist aber, daß der Punkt $(N,\hat{R})$ irgendwo zwischen den Kurven liegt. In dem Fall ziehen wir von diesem Punkt "frei" eine Kurve parallel zu den Kurven der Familie R(N) bis zum Schnittpunkt mit $R_2(N)$. Die Abzisse des Schnittpunkts ergibt dann den Wert von Z.

Den auf diese Weise bestimmten Wert von Z muß man zusammen mit N und p_{max} der gegebenen Stichprobe in die Formel (10a) oder (10b) einsetzen, den Wortschatz der Stichprobe beim Umfang N und den gefundenen Wert Z berechnen, und mit dem tatsächlichen Wortschatz $\hat{v}$ vergleichen. Alles weitere hängt sowohl von der erhaltenen Übereinstimmung als auch von den verfolgten Zielen ab.

Wenn die Divergenz zwischen $\hat{v}$ und v(N,Z) 5% nicht übersteigt, so kann man annehmen, daß die graphische Berechnung korrekt ist (man soll aber nicht vergessen, daß die Divergenz einfach durch arithmetische Fehler bei der Berechnung von v(N,Z) entstehen kann). Und wenn man nicht das Ziel einer möglichst exakten Prognose "vorwärts" oder "rückwärts" oder eines Vergleichs zweier Texte mit ähnlichen Z verfolgt, so reicht die Berechnung aus. Die Z-Werte, die wir in der vorliegenden Arbeit mit dieser "graphischen Genauigkeit" berechnet haben, sind in den Tabellen mit ~ gekennzeichnet. Dieser Fall entspricht einer 10-15%-Genauigkeit der Berechnung von Z.

Wenn man Z exakter berechnen möchte, muß man die iterative Approximation verwenden. In dieser Arbeit wurde eine vereinfachte Chordenmethode benutzt, die im Vergleich mit anderen, auf Kosten der Einfachheit einzelner Schritte schneller konvergierenden Prozeduren, eine beträchtliche Rechenökonomie bedeutet.

Wenn der aus dem Nomogramm berechnete Zipfsche Umfang als die erste Approximation betrachtet und als Z_1 bezeichnet wird, dann kann man die zweite Approximation aus der Formel

$$Z_2 = \left[\frac{\hat{v}}{v(N,Z_1)}\right]^2 \cdot Z_1$$

finden. Die dritte und die weiteren Approximationen findet man dann als

$$Z_i = Z_{i-1} + \frac{\hat{v} - v(N,Z_{i-2})}{v(N,Z_{i-1}) - v(N,Z_{i-2})} \cdot (Z_{i-1} - Z_{i-2}).$$

Nach der Berechnung eines Z_i überprüft man die Übereinstimmung zwischen der rechten und der linken Seite von (13). In der vorliegenden Arbeit wurde der Prozeß solange fortgeführt, bis die Genauigkeit für den Wortschatz nicht schlechter als 1% war. Und obwohl dies ungefähr einer 5% Genauigkeit bei der Bestimmung von Z selbst entspricht, hat eine größere Genauigkeit offensichtlich keinen Sinn, da die beobachtete zufällige Streuung der Größe Z bedeutend größer ist.

Wenn wir Z_1 graphisch berechnen, so erreichen wir diese Genauigkeit üblicherweise spätestens im dritten Schritt. Bei der Analyse einiger "exotischer " Stichproben (entweder zu großer wie z.B. {164}, oder zu kleiner {1,2,3,22,24} oder schließlich solcher mit zu großem Z {50}) lag der Punkt $(N,\hat{R})$ freilich außerhalb des Nomogramms. In dem Fall wurde der Wert Z_1 "aus der Luft gegriffen" und der Prozeß der iterativen Approximation endete beim 6-7-ten Schritt.

Als Beispiel führen wir die Berechnung von Z für einen Abschnitt aus Puškins "Kapitänstochter" {30} an. Hier ist

$$N = 5000$$

$$\hat{v} = 1671$$

$$p_{max} = 0{,}04$$

$$\hat{R} = \frac{1671}{\sqrt{5000}} = 23{,}6.$$

Der an p_{max} nächstliegende Wert auf unseren Nomogrammen ist $p_{max} = 0{,}035$. Auf dieses Nomogramm (Abb. 5) tragen wir den Punkt (5000; 23,6) auf. Wir führen "optisch" diesen Punkt entlang der konkaven Kurven bis zu der steilen Kurve $R_2(N)$, und die Stelle, wo sie geschnitten wird, projizieren wir auf die Achse N. Die erhaltene Abszisse hat ungefähr den Wert 47000. Diese Zahl betrachten wir als den Wert von Z.

Überprüfung:

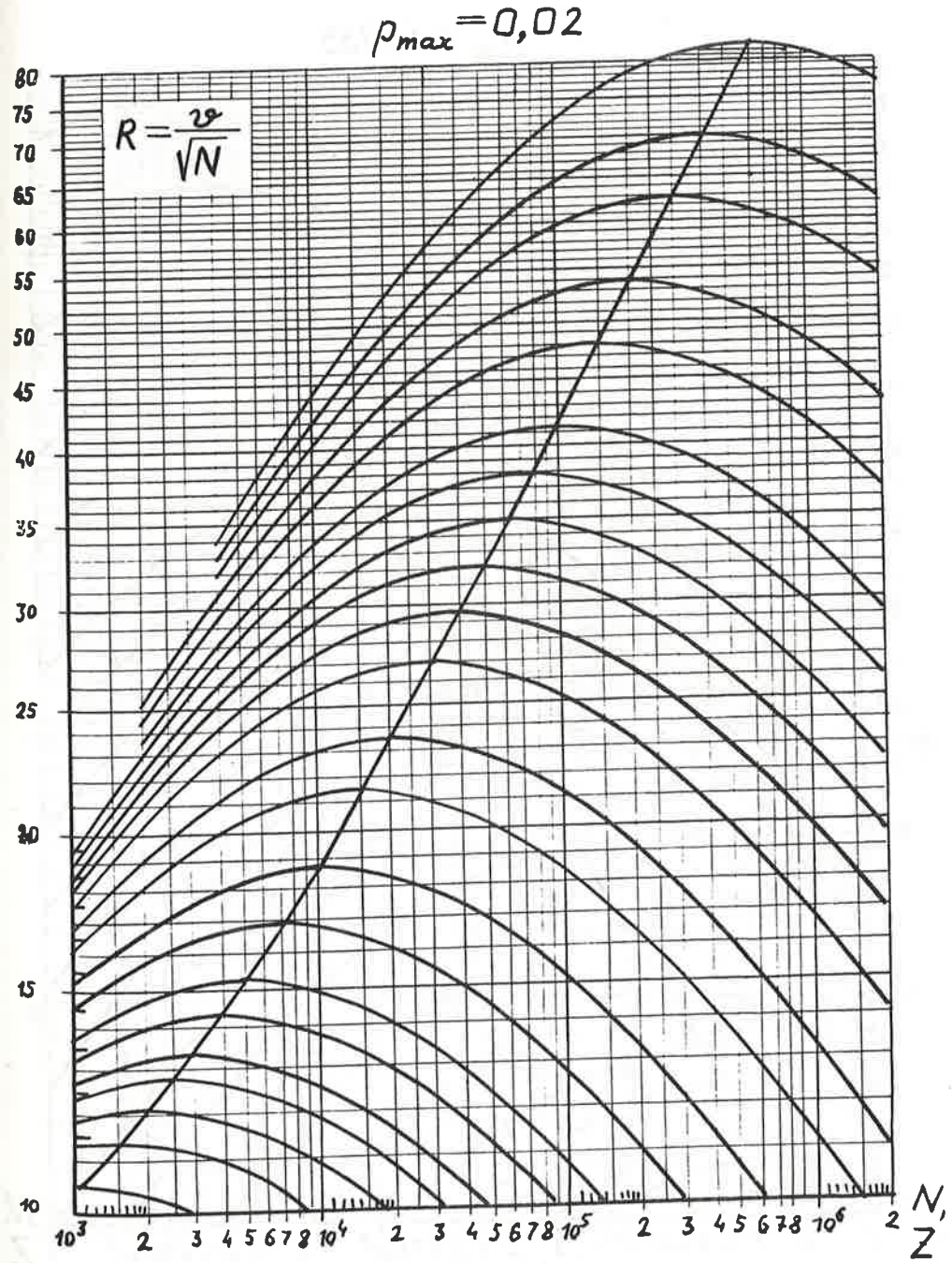
$$X = \frac{47000}{5000} = 9{,}4$$

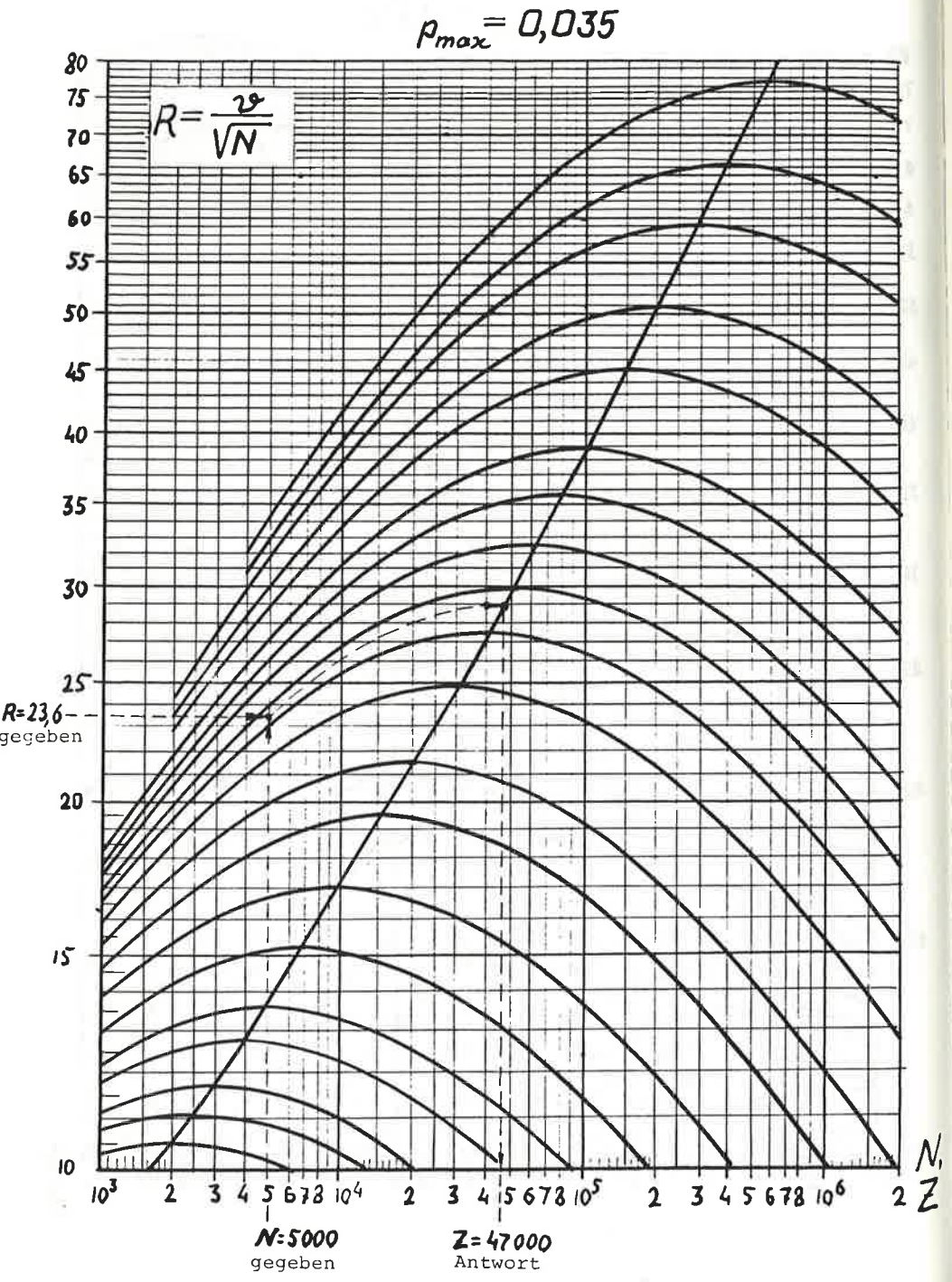
$$v(Z) = \frac{47000}{\ln\,[47000(0{,}04)]} = \frac{47000}{7{,}54} = 6234$$

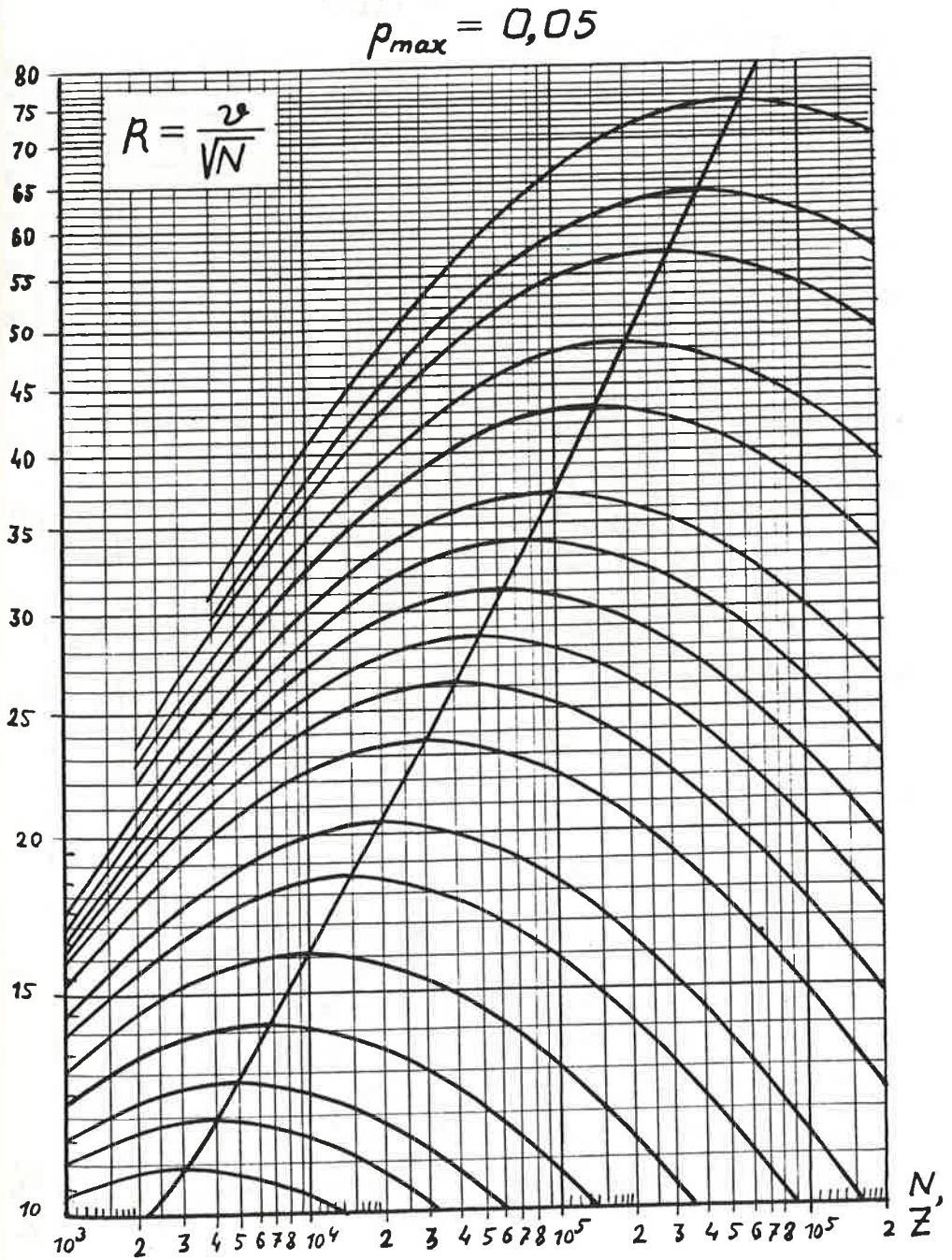
$$v(N,Z) = \frac{6234 \ln 9{,}4}{8{,}4} = 1663.$$

Da sich diese Größe von dem Wortschatz der Stichprobe (1671) um weniger als 1% unterscheidet, ist die Berechnung beendet.

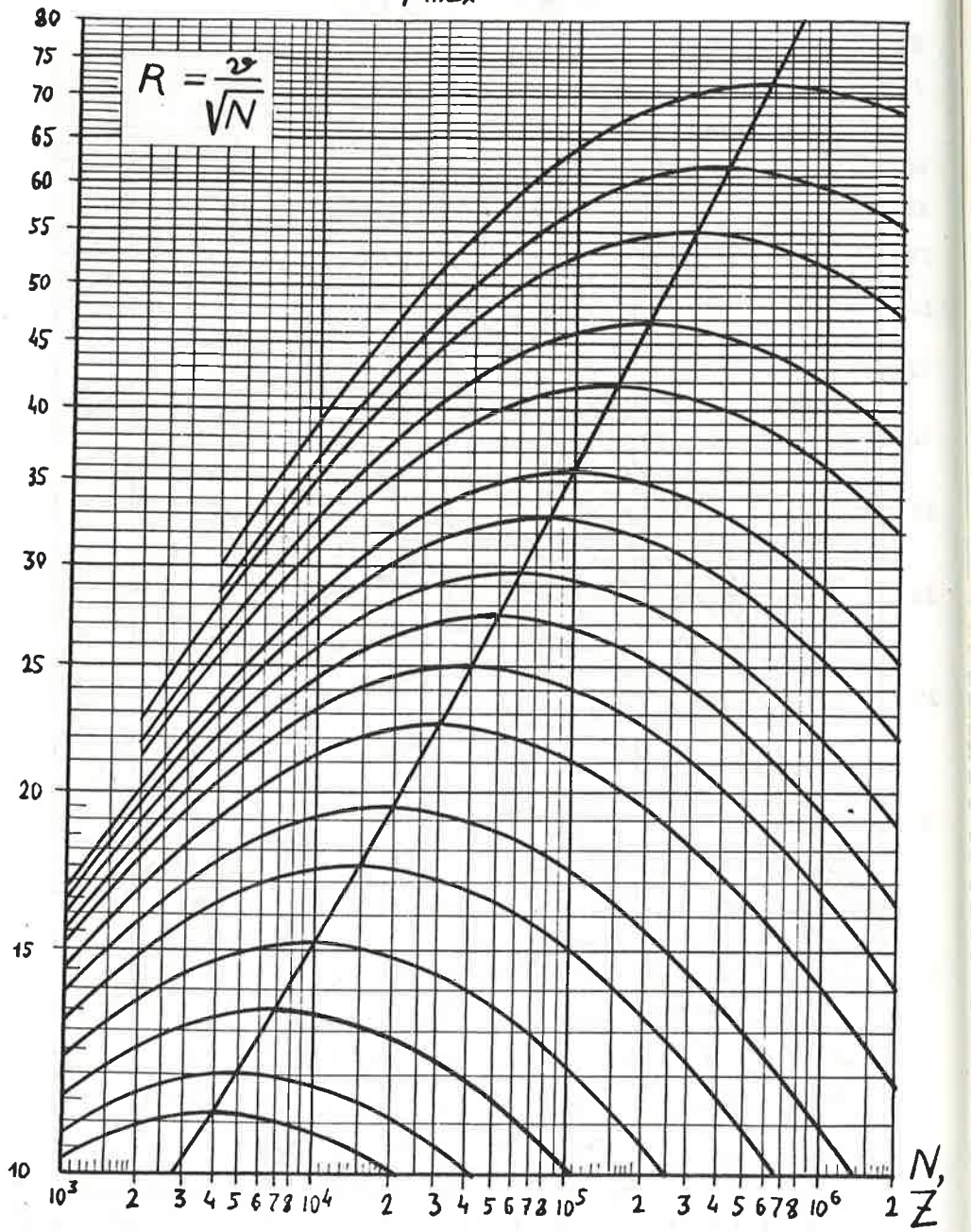
Zum Schluß möchten wir bemerken, daß es mit Hilfe angeführter Nomogramme möglich ist, den Wortschatz des Textes bei verschiedenen Umfängen vereinfacht zu berechnen. Speziell im besprochenen Beispiel kann man für N = 29345 (voller Umfang der "Kapitänstochter") $R \approx 28{,}5$ ablesen. Multipliziert man diese Größe mit der Wurzel aus der Textlänge der "Kapitänstochter", so erhält man für den gegebenen Abschnitt $28{,}5\sqrt{29345} = 4882$, was sich "nach allen Regeln" nur unbedeutend von der theoretischen Prognose 4860 unterscheidet (vgl. Zeile 2 der Tab. 3).



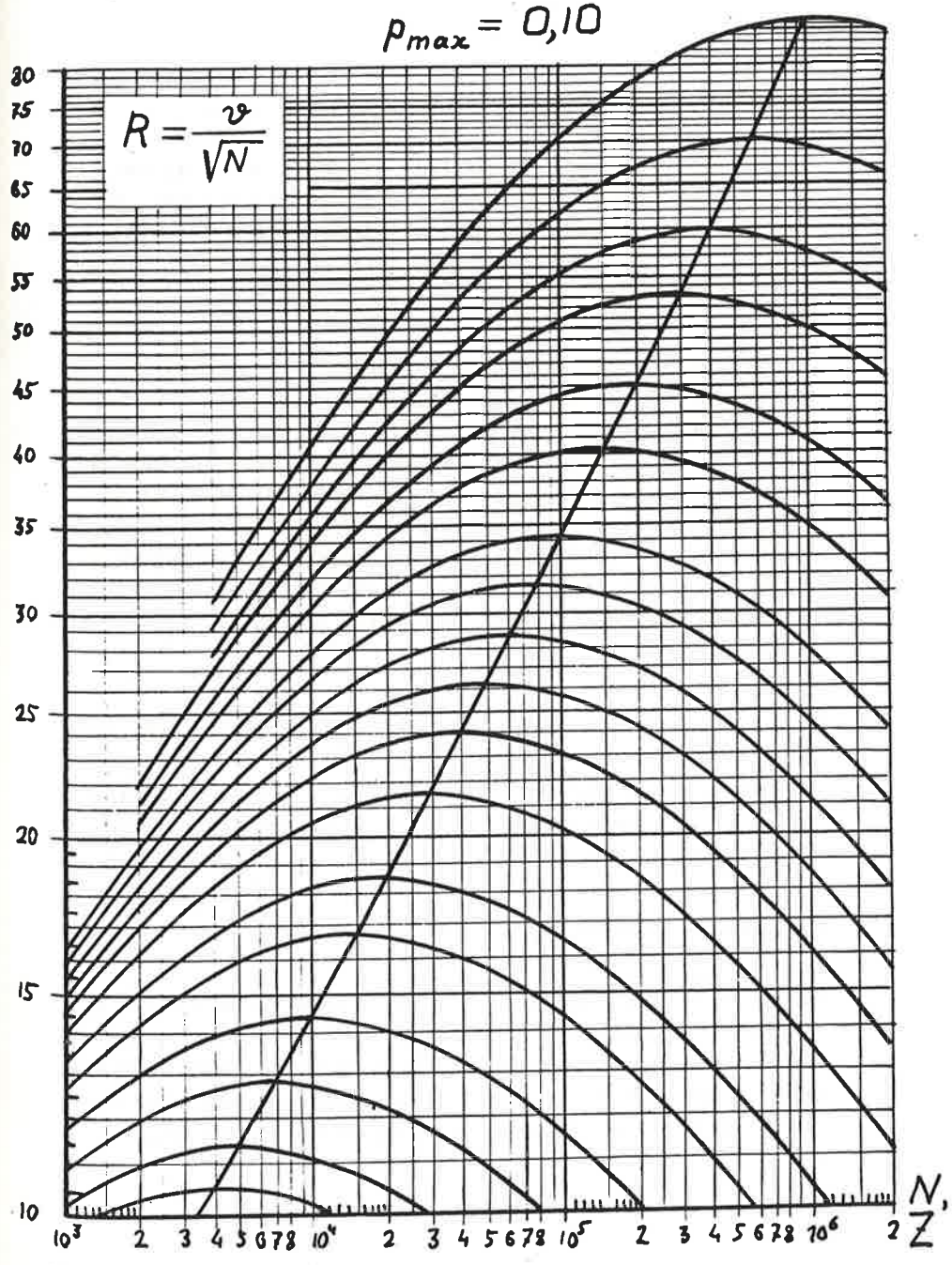

$p_{max} = 0{,}035$
$R = \frac{v}{\sqrt{N}}$
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
R=23,6
gegeben
10^3 2 3 4 5 6 7 8 10^4 2 3 4 5 6 7 8 10^5 2 3 4 5 6 7 8 10^6 2
N,
Z
N=5000
gegeben
Z=47000
Antwort


$p_{max} = 0{,}05$
$R = \frac{v}{\sqrt{N}}$
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
10^3 2 3 4 5 6 7 8 10^4 2 3 4 5 6 7 8 10^5 2 3 4 5 6 7 8 10^5 2
N,
Z

$p_{max} = 0{,}07$



$p_{max} = 0{,}10$



ANMERKUNGEN

1 Auch wenn die Meinung verbreitet ist, daß γ nicht einmal im Rahmen eines Textes konstant ist (vgl. Frumkina 1961), werden wir unten zeigen, daß dieser Effekt durch den nichtberücksichtigten Stichprobenumfang entsteht, wenn sich dieser von Z unterscheidet.

2 Genauer, wir postulieren eine solche Wahrscheinlichkeitsverteilung der Wörter in einer allgemeinen lexikalischen Gesamtheit (zufällige Stichprobe, aus der der Text generiert wird), daß bei einem Umfang N_o die mathematische Erwartung der Anzahl m-maliger Wörter durch (8) gegeben wird. Diese etwas ungewöhnliche Form der statistischen Hypothese erlaubt es, das "Raten" der unbekannten Menge der Wortwahrscheinlichkeiten zu vermeiden. Man kann eine solche Hypothese formal nicht begründen, aber die erhaltenen Resultate rechtfertigen ihre Annahme.

3 Um eine Genauigkeit von mindestens 1% zu erreichen (bei linguistischen Berechnungen besteht kein Anlass zu einer größeren Genauigkeit), reicht es, in (10a) die ersten drei Glieder zu nehmen, wenn $-0,36 \leq X - 1 \leq 0,36$, und fünf Glieder, wenn $-0,5 \leq X - 1 \leq 0,5$.

4 Vgl. auch die Formel (45b) in Orlov (1976).

5 Es ist zu betonen, daß Formel (11a) einerseits und Formeln (11b), (11c) andererseits wie auch die Formeln (40-44) in Orlov (1976) keine unterschiedlichen mathematischen Abhängigkeiten , sondern nur unterschiedliche Formen einer und derselben Abhängigkeit darstellen, ähnlich wie $(a+b)^2$ und $a^2+2ab+b^2$.

6 In den erwähnten Arbeiten ist die entsprechende Größe mit N_o bezeichnet: s. auch die Arbeit von Ju.K. Orlov. Why, How and When does the Zipf-Mandelbrot Law Fail? SMIL Quarterly, Stockholm, Skriptor, 4, 1977, 5-27.

7 Die etwas langwierigeren Methoden der approximativen Lösung der Gleichung (13) werden im Anhang 1 erläutert.

8 Wenn wir den exakten Wert der mathematischen Erwartung des Vokabularumfangs beim Stichprobenumfang N_1 kennen würden, so könnten wir durch Lösung von (13) den exakten Wert von Z für den gegebenen Text berechnen. In dem Falle würde sich die relative Standardabweichung für die Prognose ausschließlich aus δ_2 ergeben. Jedoch ist das tatsächliche Vokabular $\hat{V}^{(1)}$ im Beobachtungspunkt N_1 eine Zufallsvariable, die im Rahmen eines Konfidenzintervalls (a_1,b_1) (vgl. Abl.) schwanken kann. Dementsprechend kann auch die Prognose für den Punkt N_2 schwanken und zwar im Intervall (a_2,b_2), dessen Größe sowohl vom Intervall (a_1,b_1) als auch vom Unterschied zwischen N_1 und N_2 und von den geometrischen Eigenschaften der Kurven (10) abhängt. Wenn wir also die relative mittlere Standardabweichung S_1 im Punkt N_1 berechnet haben, müssen wir sie auf den Umfang N_2 "projizieren". Diese Aufgabe erfüllt eben der Koeffizient

$$W = \frac{b_2 - a_2}{b_1 - a_1} .$$

Der erste Summand in (15) stellt also die Dispersion der Prognose dar, die aus zufälligen Schwankungen des Vokabulars in Punkt N_1 folgen, und der zweite Summand die Dispersion des Vokabulars im prognostizierten Punkt N_2.

9 Die Zahlen in geschweiften Klammern beziehen sich auf die Numerierung der Texte und Stichproben in den Tabellen 1 - 4 (vgl. Anhang 2).

10 Eine mögliche Ursache wird unten besprochen.

11 In dieser Analyse wurden die beobachteten relativen Abweichungen (in %) algebraisch summiert.

12 Die für einige Texte charakteristische Unterbelegung des vorhandenen Vokabulars bei der Prognose "rückwärts" veranlaßte uns, die Analyse der Abweichungen in Text 3 genauer zu betrachten. Die Abweichungen in Tab. 3 sind gewissermaßen symmetrisch (während die Prognose für eine kleinere Stichprobe aus den Daten einer größeren Stichprobe höher ausfällt, so ist die umgekehrte Prognose für eine größere Stichprobe aus den Daten einer kleineren Stichprobe niedriger; bei der algebraischen Summierung der relativen Abweichungen kompensieren sich diese Fehler gegenseitig) und das Fehlen einer systematischen Verschiebung kann eben durch diese Symmetrie erklärt werden. Die Unterscheidung von "vorwärts"- und "rückwärts"-Prognosen zeigte (Tab. 3), daß in jeder Klasse der Prognosen signifikante Abweichungen vorhanden sind.

13 Wir bemerken, daß bei der Hypothese R = const., der Fehler der Prognose über den Vokabularumfang viel größer ist; für die "Kapitänstochter" (beim vollen Umfang ist R = 27,9) ergibt sich das folgende Bild:

für N = 10000, $v = 27{,}9\sqrt{10000} = 2790$
(der tatsächliche Wert ist 2432; Abweichung -12,8%);

für N = 5000, $v = 27{,}9\sqrt{5000} = 1970$
(die tatsächlichen Werte sind 1671 und 1567; Abweichungen -15,2% bzw. -20,5%).

Man vergleiche die Zeilen 4,7 und 10 in Tab. 3.

14 Offensichtlich werden die auf Abb. 2 ersichtlichen Abweichungen der experimentellen Kurve von der theoretischen durch die besonders hohe Inhomogenität der Gesamtwerke von Puškin hervorgerufen.

15 In einigen Fällen wird diese Anordnung geringfügig gestört, damit die Reihenfolge der Zeilen, die sich auf einen Text oder auf ähnliche Texte beziehen, intakt bleibt.

16 Wir bemerken, daß es bei der Untersuchung der Entsprechung zwischen Z und der Textlänge keine Möglichkeit gibt, die Unterschiede der lexikalischen Zählungen zwischen einzelnen Forschern auf einen gemeinsamen Nenner zu bringen. Bisher gibt es keine allgemein angenommene Bestimmung des Wortes und die unterschiedlichen Arten der Identifizierung der Wortform machen sich im Wortschatz bemerkbar. So schreibt z.B. Osmanov (1970): "Wir bringen absichtlich keine zusammengesetzten Verben, die aus Substantiven und Verben gebildet werden, da es im Persischen oft ziemlich schwierig ist festzustellen, ob ein Substantiv und ein Verb ein zusammengesetztes Verb oder eine Wortgruppe bilden. Wenn wir die zusammengesetzten Verben erfassen würden, so würde Unsuris Vokabular beträchtlich anwachsen". In diesem Falle würde offensichtlich auch das Z des Textes größer werden und die Übereinstimmung zwischen der Textlänge von Unsuris "Divan" {21} und seinem Zipfschen Umfang würde sich verbessern (nach Osmanovs Angaben ist X = 26400/46472 = 0,57). Infolgedessen besteht Grund zur Annahme, daß bei einer standardisierten Zählung und Benutzung der Materialien in einer Sprache die beschriebene Korrelation stärker sein müßte. (Die letztgenannten Bedingungen werden zwar in dem "Häufigkeitswörterbuch des Tschechischen" erfüllt, aber trotzdem tritt die erwartete Vergrößerung der Korrelation nicht ein, da die Angaben über die volle Länge der Texte, aus denen die Stichproben erhoben wurden (z.B. {58, 62, 64, 66} u.a.), und die Angaben über die eventuellen Gliederungen der Texte fehlen.)

17 Der Mensch bemerkt die Veränderung einer Größe, wenn sie 10-30% des ursprünglichen Wertes übersteigt (Konstanz der sogenannten differentialen Wahrnehmungsschwelle, das Weber-Fachnersche Gesetz). Kleinere Veränderungen werden von dem Menschen nicht wahrgenommen. Alle charakteristischen Abweichungen zwischen theoretischen Prognosen und tatsächlichen Beobachtungen in dieser Arbeit (nicht nur die Abweichungen des L(X) von 0.5) liegen in dem Intervall der relativen Abweichungen von ± 10-30%. Früher haben wir bemerkt (Orlov 1969a,b, 1970a, 1974, 1978b), daß bei der Prognostizierung des Vokabulars einzelner voller literarischer Texte nach Formel (7), die Benutzung der Textlänge N_o statt Z (in der überwältigenden Mehrheit der Fälle) zu einem Prognosefehler von ± 20% führt. D.h. nimmt man die Zipf-Mandelbrotsche Häufigkeitsstruktur (beschrieben durch (6), (7), (8)) als "ideales Modell" der quantitativen Organisation des literarischen Werks, dann befinden sich die beobachteten Abweichungen von diesem "Ideal" genau an der Grenze, wo der Mensch gerade anfängt eine Abweichung zu beobachten. Die Annahme der differentiellen Geschwindigkeit des Vokabularwachstums (17) als "Ideal" hat den Vorteil, daß diese Kurve universell ist und ihr kritischer Wert 0.5 für Texte beliebiger Länge konstant ist, wenn sie dem Zipf-Mandelbrotschen Gesetz folgen (d.h. wenn L(X) sich 0.5 nähert, dann kann man den Text mit (6), (7) und (8) gut beschreiben).

LITERATUR

ABULADZE, L.B., Über eine statistische Wortschatzanalyse der künstlerischen Prosa von Vaza Psavela (in georgischer Sprache). In: Šaradzenidze, T.S. (Hrsg.), Voprosy sovremennogo obščego i matematičeskogo jazykoznanija. Tbilisi, Mecniereba 1966, 94-148

A Concordance to Byron's Don Juan. Ithaca, Cornell University Press 1967

ALEKSEEV, P.M., Statističeskaja leksikografija v anglistike [Die statistische Lexikographie in der Anglistik]. In: Rozencvejg, V.Ju. (Hrsg.), Problemy prikladonoj lingvistiki, Moskva, MCPIIja 1969, 3-7

ALEKSEEV, P.M., Častotnye slovari anglijskogo jazyka i ich praktičeskoe primenenie [Häufigkeitswörterbücher des Englischen und ihre praktische Anwendung]. In: Piotrowski, R.G. (Hrsg.), Statistika reči i avtomatičeskij analiz teksta. Leningrad, Nauka 1971, 160-178

ALINEI, M. (ed.), Spogli ellettronici dell' Italiano delle origini e del duecento. Il Forme 5: Dante Alighiere. La Commedia. Bologna, Il Mulino 1971

BEKTAEV, K.B., Častotnyj slovar' jazyka Abaja [Häufigkeitswörterbuch der Sprache von Abaj]. In: Mežvuzovskaja konferencija po voprosam častotnych slovarej i avtomatizacija lingvostatičeskich rabot. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966, 36-37

BORODIN, V.V., MATER, E.A., Sostavlenie slovnikov na EVM [Die Erstellung von Wörterbüchern mit Hilfe der EDV]. In: Avtomatizacija informacionnych rabot i voprosy matematičeskoj lingvistiki. Seminar. Vypusk I. Kiev 1967, 54-71

BUDMAN, M.M., Statistika anglijskich tekstov po avtomobilestroeniju [Die Statistik englischer Texte über den Automobilbau]. In: Piotrowski, R.G. (Hrsg.), Avtomatičeskaja pererabotka teksta metodami prikladnoj lingvistiki. Kišinev, Politechničeskij Institut 1971, 274-277

BULACHOV, M.G., Materialy dlja častotnogo slovarja russkogo jazyka (Pisarev, "Realisty"). In: Leksikologija i gramatyka, Minsk 1969

FINKENSTAEDT, Th., WOLF, D., Statistische Untersuchungen des englischen Wortschatzes mit Hilfe eines Computers. Beiträge zur Linguistik und Informationsverarbeitung 16, 1969, 7-34

FRUMKINA, R.M., Statističeskie metody izučenija leksiki [Statistische Methoden der Untersuchung der Lexik]. Moskva, Nauka 1964

FRUMKINA, R.M., K voprosu o tak nazyvaemom zakone Cipfa [Zur Frage des sogenannten Zipfschen Gesetzes]. Voprosy jazykoznanija 1961, Nr. 2, 117-122

GENKEL', M.A., Častotnyj slovar' romana D.N. Mamina-Sibirjaka "Privalovskie milliony" [Häufigkeitswörterbuch des Romans "Privalovskie milliony" von D.N. Mamin-Sibirjak]. In: Meżvuzovskaja konferencija po voprosam častotnych slovarej i avtomatizacii lingvo-statističeskich rabot. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966, 34-36

GOMEZ, C.F., Vocabulario de Cervantes. Madrid 1962

GUIRAUD, P., Les caractères statistiques du vocabulaire. Paris, Presses universitaires de France 1954

HART, C., A concordance to "Finnegans Wake". Minneapolis, University of Minnesota Press 1963

IMNAJŠVILI, I., Konkordanz zum georgischen Evangelium (in georgischer Sprache). Tbilisi, AN GSSR 1948

JELÍNEK, J., BEČKA, J.V., TEŠITELOVÁ, M., Frekvence slov, slovních druhů a tvarů v českém jazyce [Häufigkeit der Wörter, Wortarten und Wortformen im Čechischen]. Prag, Státni Pedagogické Nakladetelstvi 1961

JOSSELSON, H.H., The Russian word count and frequency analysis of grammar categories of standard literary Russian. Detroit, University Press 1953

KALININ, V.M., Nekotorye statističeskie zakony matematičeskoj lingvistiki [Einige statistische Gesetze der mathematischen Linguistik]. Problemy kibernetiki 11, 1964, 245-255

KALININ, V.M., Funkcionaly, svjazannye s raspredeleniem Puassona, i statističeskaja struktura teksta [Die mit der Poisson-Verteilung zusammenhängenden Funktionale und die statistische Struktur des Textes]. Trudy matematičeskogo instituta imeni V.A. Steklova 29, 1965, 182-197

KALININA, E.A., Častotnyj slovar' russkogo pod'jazyka elektroniki [Häufigkeitswörterbuch der russischen Teilsprache der Elektronik]. In: Piotrowski, R.G. (Hrsg.), Statistika reči. Leningrad, Nauka 1968, 144-150

KOLGUŠKIN, A.N., Lingvistika v voennom dele [Die Linguistik im Militärwesen]. Moskva, Voenizdat 1970

KUNICKIJ, V.N., Jazyk i slog komedii A.S. Griboedova "Gore ot uma" [Sprache und Stil von Griboedovs Komödie "Verstand schafft Leiden"]. Kiev 1896

KVARACCHELIJA, G.S., Opyt sravmitel'no-statističeskogo analiza leksiki statističeskimi metodoni [Versuch einer vergleichend-statistischen Analyse des Lexikons mit statistischen Methoden]. Voprosy sovremennogo obščego i matematičeskogo jazykoznanija, Vol. 2 Tbilisi, Mecniereba 1966, 152-182

LEVITSKIJ, V.V., Častotnyj slovar' učebnych posobij med-instituta [Häufigkeitswörterbuch der Lehrwerke des medizinischen Instituts]. Moskva, Izdatel'stvo Moskovskogo Universiteta 1966

LJATINA, A.M., Opyt statističeskogo analiza jazyka pisatelja (po materialam "Podnjatoj celiny" Šolochova) [Versuch einer statistischen Analyse der Schriftstellersprache (am Beispiel von Šolochovs "Neuland unterm Pflug")].Dissertation. Leningrad 1968

LUDIN, D.M., Opyt opisanija statističeskimi metodami sovremennogo afganskogo jazyka (puštu) [Versuch einer Beschreibung des modernen Afghanischen (Puschtu) mit Hilfe statistischer Methoden]. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1971

MANDELBROT, B., An information theory of the statistical structure of language. In: Jackson, W. (Hrsg.), Communication Theory, New York, Academic Press 1953, 503-512

MANDELBROT, B., O rekurrentnom kodirovanii, ograničivajuščem vlijanie pomech [On recurrent noise limiting coding]. In: Siforov, V.I. (Hrsg.), Teorija peredaci soobscenij. Moskva, Izdatel'stvo inostrannoj literatury 1957, 139-157

MUSSO, N., Le vocabulaire de Figaro dans „Le mariage". Etudes de linguistique appliquée 6, 1972, 89-98

NADAREJŠVILI, I.S., ORLOV, Ju.K., Ob upotreblenii slov različnoj častosti v poeme Rustaveli "Vitjaz' v tigrovoj škure" [Über die Verwendung von Wörtern unterschiedlicher Häufigkeit in Rustavelis Poem "Der Held im Tigerfell"]. Soobščenija AN GSSR 55, 1969 (Nr. 2), 505-508

NADAREJŠVILI, I.S., ORLOV, Ju.K., Rost leksiki kak funkcija dliny teksta (na primere proizvedenij L.N. Tolstogo i Dž. Džojsa) [Das Anwachsen der Lexik als Funktion der Textlänge (am Beispiel der Werke von L.N. Tolstoj und J. Joyce)]. Soobščenija AN GSSR 64, 1971 (Nr. 3), 549-552

NOVAK, D.A., Častotnyj slovar' sovremennogo moldavskogo i rumynskogo jazykov [Häufigkeitswörterbuch der moldauischen und der rumänischen Gegenwartssprache]. Dissertation, Leningrad 1962

ORLOV, Ju.K., Obobščenie zakona Cipfa [Eine Verallgemeinerung des Zipfschen Gesetzes]. In Rozencvejg, V.Ju. (Hrsg.), Problemy prikladnoj lingvistiki. Moskva, MGPIIJa 1969a, 255-261

ORLOV, Ju.K., Leksičeskij spektr literaturnych tekstov i dinamika rosta slovarja [Das lexikalische Spektrum literarischer Texte und die Dynamik des Vokabularzuwachses]. In: Rozencvejg, V.Ju. (Hrsg.), Problemy prikladnoj lingvistiki. Moskva, MGPIIJa 1969b, 250-254

ORLOV, Ju.K., O statističeskoj strukture soobščenij, optimal'nych dlja čelovečeskogo vosprijatija [Zur statistischen Struktur der für die menschliche Perzeption optimalen Nachrichten]. Naučno-techničeskaja informacija Serija 2, 1970a (Nr. 8), 11-16

ORLOV, Ju.K., Obobščenie zakona Cipfa-Mandel'brota [Eine Verallgemeinerung des Zipf-Mandelrotschen Gesetzes]. Soobščenija AN GSSR 57, 1970b (Nr. 1), 37-40

ORLOV, Ju.K., Častotnye struktury konečnych soobščenij v nekotorych estestvennych informacionnych sistemach [Die Häufigkeitsstrukturen endlicher Nachrichten in einigen natürlichen Informationssystemen]. Habilitationsschrift, Tbilisi 1974

ORLOV, Ju.K., Obobščennyj zakon Cipfa-Mandel'brota i častotnye struktury informacionnych edinic različnych urovnej [Das verallgemeinerte Zipf-Mandelbrotsche Gesetz und die Häufigkeitsstrukturen der Informationseinheiten verschiedener Ebenen]. In: Guseva, E.K. (Hrsg.), Vyčislitel'naja lingvistika. Moskva, Nauka 1976, 179-202

ORLOV, Ju.K., Statističeskoe modelirovanie rečevych potokov [Die statistische Modellierung des Redeflusses]. In: Piotrowski, R.G. (Hrsg.), Voprosy kibernetiki 41: Statistika reči i avtomatičeskij analiz teksta. 1978 , 66-106

OSMANOV, M.N.O., Častotnyj slovar' Unsuri [Häufigkeitswörterbuch von Unsuri]. Moskva, Nauka 1970

OVSIENKO, Ju.G., Slovar' russkoj razgovornoj reči [Wörterbuch der russischen Umgangssprache]. In: Mežvuzovskaja konferencija po voprosam častotnych slovarej i avtomatizacija lingvostatističeskich rabot. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966, 23-25

SAMBOR, J., Badania statystyczne nad slownictwem (na materiale „Pana Tadeuza"). Wroclaw - Warszawa - Krakow, 1969

SÉNÈQUE, De Clementia. Index verborum. Relèves statistiques. La Haye, Mouton 1968

SÉNÈQUE, De Brevitate Vitae. Index verborum. Relèves statistiques. La Haye, Mouton 1968

SINENKO, G.D., Častotnyj slovar' "Povesti o detstve" F.V. Gladkova i voprosy leteraturnogo stilja [Häufigkeitswörterbuch von "Erzählungen von der Kindheit" von F.V. Gladkov und Probleme des literarischen Stils]. Vesnik Belaruskaga dzjaržajnaga universiteta imja V.I. Lenina. Seryja IV/1. Minsk 1973, 40-46

ŠTEJNFELD, E.A., Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka [Häufigkeitswörterbuch der modernen russischen Schriftsprache]. Tallin, NII Pedagogiki ESSR 1963

SUDAVIČENE, L., Iz opyta sostavlenija častotnogo slovarja jazyka K. Paustovskogo [Aus dem Versuch der Erstellung des Häufigkeitswörterbuchs der Sprache von K. Paustovskij]. Vil'njus, Učenye zapiski vysšich učebnych zavedenij Litovskoj SSR. Jazykoznanije 22, 1971

TEŠITELOVÁ, M., On the so-called vocabulary richness. Prague Studies in Mathematical Linguistics 3, 1972, 103-120

TOMBEUR, P., R. de Saint-Trond: Gesta Abbatum trudonensium I-VIII. Index verborum. Relèves statistiques. La Haye, Mouton 1965

TVOROGOV, O.V., O primenenii častotnych slovarej v istoričeskoj leksikologii russkogo jazyka [Über die Anwendung von Häufigkeitswörterbüchern in der historischen Lexikologie des Russischen]. Voprosy jazykoznanija 1967/2, 109-117

VJALKINA, L.V., LUKINA, G.N., Opyt primenenija nekotorych metodovmatematičeskoj statistiki k izučeniju drevnerusskich tekstov [Versuch der Anwendung einiger mathematisch-statistischer Methoden zur Untersuchung altrussischer Texte]. In: Avanesov, R.I. (Hrsg.), Issledovanija po istoričeskoj leksikologii drevnerusskogo jazyka. Moskva, Nauka 1964, 298-307

YULE, G.U., The statistical study of literary vocabulary. London, Cambridge University Press 1944

ŽANTIEVA, D.G., Džejms Džojs [James Joyce]. Moskva, Vyssaja škola 1967

ZASORINA, L.N., Avtomatizacija i statistika v leksikografi [Automatisierung und Statistik in der Lexikographie]. Leningrad, Izdatel'stvo Leningradskogo Universiteta 1966

ZIPF, G.K., The psycho-biology of language. Boston, Houghton Mifflin Company 1935

ZIPF, G.K., Human behavior and the principle of least effort. Cambridge (Mass.), Addison-Wesley Press 1949

Eine Untersuchung statistischer Gesetzmässigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen

Ju.K. Krylov, Moskau

1. Einführung

Wir werden die Lexik einer natürlichen Sprache als Makrosystem betrachten, das aus einer sehr großen Zahl von Elementen - den lexikalischen Einheiten - besteht. Die Anzahl dieser Elemente beläuft sich auf Zehntausende, ist also groß genug, um die Anwendung statistischer Methoden für ihre Beschreibung möglich zu machen.

Gegenwärtig gibt es nicht wenige Arbeiten, die sich mit der Untersuchung des sprachlichen Wortschatzes mit statistischen Mitteln beschäftigen (vgl. z.B. Zipf 1949; Piotrowski 1968; Arapov/Efimova/Šrejder 1975). In den weitaus meisten Fällen jedoch wird die Lexik auf der syntagmatischen Ebene betrachtet, d.h. auf der Ebene ihrer Funktion in der Rede. Gleichzeitig stimmen bekanntlich (vgl. Andreev 1967: 17) die syntagmatischen Wahrscheinlichkeiten bei weitem nicht immer mit den paradigmatischen überein. Im Unterschied zu den erwähnten Arbeiten wird in der vorliegenden der lexikalische Bestand der Sprache namentlich auf der paradigmatischen Ebene untersucht, d.h. aufgrund der Betrachtung eines Grundwortschatzes nach den Daten aus Definitionswörterbüchern.

In jedem Wörterbucheintrag findet sich eine Anzahl verschiedener Bedeutungen, die das jeweilige Wort besitzt. In einer natürlichen Sprache ist also jedem Wort eine bestimmte natürliche Zahl i zugeordnet, die Anzahl seiner Bedeutungen, die ein objektives Maß für den semantischen Gehalt dieses Wortes darstellt. Dabei ist die Anzahl aller möglichen Bedeutungen unbegrenzt, $i \in \{1,2,\ldots,\infty\}$, da in der Sprache im Prinzip für eine beliebige vorgegebene Zahl die Bildung eines Wortes möglich ist, dessen Bedeutungszahl diese übersteigt. (Wenn wir annehmen, daß ein bestimmtes Wort zu einem festen Zeitpunkt i Bedeutungen besitzt,

können wir nicht ausschließen, daß es nach Ablauf einer gewissen Zeit eine zusätzliche Bedeutung bekommt, und seine Bedeutungszahl i+1 wird.)

Die Bedeutungszahl i ist ein Maß für den semantischen Gehalt eines einzelnen Wortes. Die Lexik insgesamt soll durch den prozentualen Anteil an Wörtern pro Bedeutungszahl charakterisiert werden, d.h. durch die Wahrscheinlichkeiten P_i dafür, daß ein zufällig gewähltes Wort i-deutig ($i = 1,2,\ldots,\infty$) ist. Mit anderen Worten ist für eine beliebige Sprache die Verteilung der Zufallsvariablen I - der Bedeutungszahl eines einzelnen Wortes - ein objektives Maß für die semantische Struktur ihrer Lexik insgesamt. Im Folgenden werden wir sie die paradigmatische Wahrscheinlichkeitsfunktion oder kürzer paradigmatische Funktion nennen.

Zwei prinzipielle Fragen stellen sich:

1. Gibt es ein derartiges für alle natürlichen Sprachen universelles Gesetz oder ist es für die verschiedenen lexikalischen Systeme der einzelnen Sprachen spezifisch?

2. Ist es mit unserem heutigen Wissen möglich, die konkrete Form der paradigmatischen Wahrscheinlichkeitsfunktion zu bestimmen?

Die vorliegende Arbeit stellt einen Versuch dar, eben diese Fragen zu beantworten. Wir verweilen zunächst bei der zweiten Frage und verschieben die Antwort auf die erste auf Kapitel 3.

2. Die empirische Wahrscheinlichkeitsfunktion mehrdeutiger Wörter nach ihrer Bedeutungszahl

Natürlich ist die Aufstellung einer empirischen Wahrscheinlichkeitsfunktion nur aufgrund der Daten von Definitionswörterbüchern möglich. Nun sind bekanntlich die Methoden zur Abgrenzung einzelner Wortbedeutungen noch sehr unzureichend, so daß die verschiedenen Definitionswörterbücher in den Bedeutungszahlen der jeweiligen Wörter häufig voneinander abweichen. Betrachten wir dies etwas ausführlicher.

Alle Fehler kann man in zufällige und systematische unterteilen. Unter zufälligen Fehlern wollen wir solche verstehen, bei denen die Abweichung in positiver und negativer Richtung ungefähr gleichwahrscheinlich ist. Derartige Fehler können die Gestalt der Wahrscheinlichkeitsfunktion kaum wesentlich verändern, weil in jeder Klasse die Anzahl der Wörter die aufgrund dieser Fehler fortfallen, durch die Zahl derjenigen kompensiert wird, die ebenso fälschlich hinzugezählt werden. Das Auftreten solcher Fehler kann also lediglich zu einer größeren Streuung führen.

Wesentlich ernster stellen sich die systematischen Fehler dar, die durch prinzipielle methodische Fehler bei der Erstellung der Wörterbücher oder bei der Abgrenzung der einzelnen Bedeutungen entstehen. Dabei ist es möglich, daß für eine ganze, umfangreiche Wortgruppe die zugehörigen Bedeutungszahlen immer in die gleiche Richtung verfälscht werden, was zu einer wesentlichen Abweichung der empirischen Wahrscheinlichkeitsfunktion von der sprachlichen Realität führen kann. Ein Beispiel eines solchen systematischen Fehlers wäre die Über- oder Unterrepräsentation von speziellem und terminologischem Vokabular in einem allgemeinsprachlichen Wörterbuch, was entsprechend einen erhöhten oder verminderten prozentualen Gehalt von ein- und zweideutigen Wörtern zur Folge hätte. Auf diese Weise können sich auch, wenn sich die Verfasser verschiedener Definitionswörterbücher verschiedener Methoden bedienen, die aufgrund dieser Wörterbücher aufgestellten Verteilungen stark voneinander unterscheiden.

Daher muß an erster Stelle untersucht werden, ob den zur Zeit vorhandenen Wörterbüchern gleichartige oder unterschiedliche Wahrscheinlichkeitsfunktionen entsprechen. Im Fall einer einheitlichen Wahrscheinlichkeitsfunktion können wir offensichtlich damit rechnen, daß sie die Situation in gewissem Maße so wiederspiegelt, wie sie in der Sprache tatsächlich herrscht. Liegen mehrere Verteilungen vor, ist diejenige als besser anzusehen, die näher an der aus rein theoretischen Überlegungen gewonnenen Wahrscheinlichkeitsfunktion liegt - vorausgesetzt wir verfügen über eine solche.

Eine experimentelle Untersuchung des Verteilungsgesetzes für die Vorkommenshäufigkeit von Wörtern bei einer bestimmten Anzahl

von Lesarten als Zufallsvariable wurde anhand der Definitionswörterbücher MAS (vierbändig), SSRLJa (17-bändig) und des Wörterbuchs von Ožegov (WO) durchgeführt. Die Resultate, die sich für die Verben ergaben, sind in Tabelle 1 dargestellt. Wegen der Stichprobenrepräsentativität wurden der Bearbeitung alle Verben unterzogen, die in wenigstens einem der Wörterbücher (s. 2. Spalte) unter dem entsprechenden Buchstaben (äußerst linke Spalte) aufgeführt sind. Weiter ist in jeder Spalte die Anzahl m_i der i-deutigen Wörter eingetragen, in der Spalte mit der Bezeichnung "$m_i \geq 10$" die Summe aller Fälle mit 10 und mehr Lesarten. In den Zeilen hinter dem Summensymbol "Σ" stehen die Daten für das MAS und für das WO für die Buchstaben A-I,S und T zusammen. Schließlich sind in den beiden untersten Zeilen die relativen Häufigkeiten ν_i von i-deutigen Verben in den Wörterbüchern angegeben. Diese Häufigkeiten sind bekanntlich (vgl. Dunin-Barkovskij/Smirnov 1955) effiziente und unverzerrte Schätzungen für die Wahrscheinlichkeiten P_i und bestimmen so die Form der empirischen paradigmatischen Verteilung zu dem entsprechenden Wörterbuch.

Leider mußte die Untersuchung wegen des großen Arbeitsaufwands auf den Wortbestand unter den einzelnen Buchstaben beschränkt werden. Wie aber weiter unten gezeigt wird, reicht der Umfang der durchgeführten Untersuchungen bereits aus, um auf die Verteilung der Verben als Funktion der Bedeutungszahl für das Lexikon der russischen Sprache insgesamt zu schließen.

Zur Bestimmung der Bedeutungszahl eines Wortes wurde folgendermaßen verfahren: Um sicherzustellen, daß ein synchroner Ausschnitt der Sprache entsteht, wurden Wörter mit den Vermerken "veraltet", "arch." und "folklor." (wie z.B. sovivat'; ds. wie svivat') nicht in Betracht gezogen. Auch Wörter mit dem Zusatz "landschaftl." wurden ausgesondert. Von den Varianten ein und desselben Wortes (z.B. strogat', strugat') wurde nur eine gezählt. In Übereinstimmung mit der Auffassung von Bondarko (1967) wurden Aspektpaare, die mit Hilfe von Suffixen gebildet werden, als Varianten eines Verbs betrachtet; mit Hilfe von Präfixen gebildete Aspektpartner wurden als zwei eigenständige Lexeme angesehen. Schließlich wurden Wörter, die den Vermerk "Pass. zu" tragen, als

Tabelle 1. Verteilung mehrdeutiger Verben mit unterschiedlichen Anfangsbuchstaben

		I	2	3	4	5	6	7	8	9	$m_i \geq 10$	x^2	x^2_r	s
A	MAS	48	19	2	0	0	0	0	0	0	0	2,2	3,84	2
	WO	47	9	2	0	0	0	0	0	0	0			
B	MAS	96	47	13	5	2	I	0	0	0	2	4,9	5,99	3
	WO	119	33	13	2	I	4	0	0	0	2			
V	MAS	875	301	115	29	14	10	3	1	0	2	7,5	12,6	7
	WO	539	202	55	28	12	6	1	2	1	1			
G	MAS	115	43	13	1	4	1	1	0	1	0	0,1	7,81	4
	WO	85	31	3	2	0	0	1	0	0	0			
D	MAS	464	129	25	10	13	5	0	1	1	2	8,1	9,49	5
	WO	157	61	12	8	3	2	4	0	0	0			
E	MAS	12	4	3	3	0	0	0	0	0	0	0,4	3,84	2
	WO	8	1	0	1	1	1	0	0	0	0			
Ž	MAS	40	8	5	3	1	0	0	0	0	0	5,6	5,99	3
	WO	10	7	3	1	1	0	0	0	1	0			
I	MAS	461	154	40	6	2	0	0	0	0	0	3,7	5,99	3
	WO	193	59	8	1	1	0	0	0	0	0	9,1	7,81	4
	SSRLJa	428	152	45	11	5	4	2	2	0	0			
S	MAS	779	394	145	89	29	25	14	8	12	18	20,3	15,5	9
	WO	467	208	61	28	13	6	5	4	0	8	31,7	16,9	10
	SSRLJa	739	415	153	93	67	42	25	20	21	51			
T	MAS	189	88	40	24	11	4	3	3	2	1	1,7	11,1	6
	WO	113	49	19	16	5	2	1	1	0	1			
Σ	MAS	3079	1187	401	170	76	46	20	13	16	25	7,5	18,3	11
	WO	1730	660	182	88	39	21	12	7	2	12			
ν_i	MAS	0,612	0,240	0,08	0,034	0,015	0,0091	0,004	0,0028					
	WO	0,626	0,239	0,07	0,032	0,014	0,0075	0,004	0,0025					

selbständige Einheiten aufgenommen. Wenn ein Verb in einem Lexikoneintrag neben einer Anzahl von Lesarten zusätzlich den Vermerk "Pass. zu" hatte, wurde die Anzahl der gemeinsamen Bedeutungen hinzugezählt. Für das Verb soskrebat'sja z.B. gibt das SSRLJa an: "Pass. zu soskrebat'". Zu soskrebat' werden zwei Lesarten angeführt: "1. Durch Schaben reinigen, an der Oberfläche säubern; 2. Übertr., ugs. Mit Mühe (einen Geldbetrag) zusammenbringen." Nach den oben erläuterten Verfahren wurden die Verben soskrebat' und soskrebat'sja als zwei selbständige Einheiten mit je zwei Bedeutungen in die Stichprobe aufgenommen. Das gleiche Wörterbuch führt für das Verb sprašivat'sja fünf Lesarten an und hat als Nr. 6: "Pass. zu sprašivat' (in den Bedeutungen 1-3)", wobei das Verb sprašivat' ebenfalls fünf Lesarten aufweist. Demnach wurden die Verben sprašivat' und sprašivat'sja als zwei Lexeme mit fünf bzw. acht Lesarten aufgenommen.

Um nun zur Erörterung der Resultate zu kommen, wollen wir zunächst zeigen, daß für alle drei Wörterbücher die absolute Zahl der i-deutigen Wörter in Abhängigkeit von i monoton fällt (s. Tab. 1), und die Anzahl der mehrdeutigen Wörter ($i \geq 2$) ungefähr die Hälfte des Wortbestandes ausmacht.

Wir beginnen die Analyse der empirischen Verteilung mit einer Gegenüberstellung von WO und MAS und betrachten die Zeilen der Tabelle 2, die die angegebenen Wörterbücher insgesamt charakterisieren (Symbole Σ und ν_i). Abgesehen von dem beträchtlichen Unterschied in der absoluten Anzahl der i-deutigen Wörter, der lediglich den jeweils verschiedenen Umfang der betrachteten Wörterbücher reflektiert, ist offensichtlich die Art der Veränderung der Vorkommenshäufigkeit ν_i i-deutiger Wörter gleich. Damit können wir die Hypothese aufstellen, daß die paradigmatischen Verteilungen dieser Wörterbücher identisch sind.

In der Sprache der Statistik bedeutet das, daß die erhobenen Stichproben aus ein und derselben Grundgesamtheit stammen. Bekanntlich (vgl. van der Waerden 1960) wird eine Hypothese über die Zugehörigkeit zweier Stichproben zu einer Grundgesamtheit mit Hilfe des sogenannten "Chi-Quadrat"-Tests überprüft. Dabei berechnet sich der Ausdruck X^2 folgendermaßen:

Tabelle 2. Zahl der Verben mit gegebenen Anfangsbuchstaben in MAS, WO und SSRLJa

	A	B	V	G	D	E	Ž	I	S	T	n	X^2	χ^2_r
SSRLJa								650	1626		2276		
MAS	69	164	1350	179	650	22	56	663	1513	365	5031	104,5	16,9
WO	58	172	848	131	250	12	23	262	800	207	2763		
MAS/WO	0,84	1,05	0,63	0,73	0,38	0,54	0,41	0,40	0,53	0,57	0,575		

$$X^2 = n_1 n_2 \sum_{i=1}^{s} \frac{1}{m_i' + m_i''} \left(\frac{m_i'}{n_1} - \frac{m_i''}{n_2} \right)^2 \qquad (2.1)$$

wobei m_i' und m_i'' die Anzahl der i-deutigen Wörter im ersten bzw. zweiten Wörterbuch bedeuten, $n_1 = \sum_{i=1}^{s} m_i'$ und $n_2 = \sum_{i=1}^{s} m_i''$. Der X^2-Wert muß nun mit der theoretischen Größe χ^2_r verglichen werden, welche für eine bestimmte Anzahl von Freiheitsgraden (in unserem Fall r = s - 1) und für das Signifikanzneveau α aus speziellen Tabellen entnommen werden kann (vgl. Dunin-Barkovskij/Smirnov 1955). Ist $X^2 < \chi^2_r$, wird die Hypothese als mit den experimentellen Ergebnissen verträglich angesehen. Andernfalls muß sie zugunsten einer Alternativhypothese ("die Verteilungen sind verschieden") abgelehnt werden.

Die Berechnung der X^2-Werte nach (2.1) ergab $X^2 = 7,5 < \chi^2_{10} = 18,3$ (hier und im Folgenden werden alle Berechnungen auf dem Signifikanzniveau α = 0,05 durchgeführt, wie es in der mathematischen Statistik für technische Fragestellungen üblich ist). Somit liegt für alle Verben, die in WO und MAS angeführt sind, die gleiche Charakteristische paradigmatische Verteilung vor.

Aus den Daten in Tab. 1 geht außerdem noch hervor, daß sich die empirischen Verteilungen der Wörter mit verschiedenen Anfangsbuchstaben sogar innerhalb ein und desselben Wörterbuchs wesentlich unterscheiden. So ist z.B. der Prozentanteil mehrdeutiger Wörter mit den Anfangsbuchstaben A und I bedeutend geringer als für das Gesamtwörterbuch. Dieser Umstand rührt offensichtlich von der Fluktuation des Anteils fremdsprachlicher, spezieller und terminologischer Lexik

unter den verschiedenen Buchstaben her. Andererseits ist die Anzahl der Lesarten eines bestimmten Wortes umso größer, je länger es bereits in der Sprache fungiert (vgl. Arapov/Cherz 1974). Daher schlägt sich auch die Fluktuation des Wortbestands unter verschiedenen Anfangsbuchstaben in spezifischen paradigmatischen Verteilungen, die diesen Buchstaben entsprechen, nieder.

Nichtsdestoweniger korrelieren augenscheinlich in WO und MAS die Abweichungen der Verteilungen von der gesamtsprachlichen Verteilung miteinander (s. Tab.1). In der Tabelle sind in der Spalte "X^2" die nach (2.1) berechneten Werte für die Wörter mit den entsprechenden Anfangsbuchstaben angegeben.

Wegen der unterschiedlichen Wortanzahl wurden die Berechnungen mit unterschiedlicher Klassenzahl s, die jeweils in der äußerst rechten Spalte angegeben ist, durchgeführt. Die Klassenzahl s wurde so gewählt, daß die Anzahl der Einheiten in keiner der Klassen kleiner als 5 wurde, weil sonst das Resultat des χ^2-Tests erheblich verzerrt werden könnte (Dunin-Barkovskij/Smirnov 1955). In die höchste Klasse wurden auch die Verben aufgenommen, deren Bedeutungszahl s überstieg, so daß sich also in der s-ten Klasse alle Wörter mit einer Bedeutungszahl i ≥ s befinden. In Tab. 1 ist zum Beispiel für WO die Anzahl der Wörter mit dem Anfangsbuchstaben D mit einer Bedeutungszahl von 1 bis 4 größer als 5, während die Zahl der fünfdeutigen Wörter nur noch drei beträgt. Somit ist s = 5, und dieser Klasse werden 22 Wörter für MAS und 9 für WO zugerechnet.

Der theoretische Wert χ^2_r mit r = s - 1 Freiheitsgraden steht in der vorletzten Spalte von Tab. 1. Die Gegenüberstellung der experimentellen und der theoretischen χ^2-Werte gab bis auf einen Fall (beim Buchstaben S) Anlaß dazu, die Verteilungen für die verglichenen Wörterbücher als verschieden anzusehen. Nur im Fall des Buchstabens S fällt die paradigmatische Funktion für MAS viel langsamer als die für WO. Aufgrund dieser Ergebnisse können wir also feststellen, daß die empirischen Verteilungen der Wörterbücher (sowohl insgesamt wie auch nach Anfangsbuchstaben getrennt betrachtet) zusammenfallen.

Letzteres ist besonders interessant, wenn man in Betracht zieht, daß sich die Wörterbücher in der Anzahl der erfaßten Wörter unter

den verschiedenen Buchstaben wesentlich voneinander unterscheiden. Entsprechende Daten finden sich in Tab. 2.

Bei der Erstellung eines allgemeinsprachlichen Wörterbuchs ergibt sich, unabhängig vom Umfang, die natürliche Forderung, daß sich die Anzahl der Wörter, die unter einem bestimmten Buchstaben angeführt sind, in demselben prozentualen Verhältnis zu den anderen befindet wie in der ganzen Sprache. Also dürften sich die jeweiligen Anteile von Wörtern unter einem bestimmten Anfangsbuchstaben in den Wörterbüchern nicht merklich unterscheiden.

Diese Forderung ist für MAS und WO nicht erfüllt. Tab. 2 zeigt die jeweilige Verbanzahl nach Anfangsbuchstaben für beide Wörterbücher. Auch ohne statistische Berarbeitung fällt der krasse Unterschied ins Auge. Beim Buchstaben B ist das WO, das im Durchschnitt den halben Umfang von MAS aufweist, sogar reicher als dieses unter demselben Buchstaben. So verwundert es nicht, daß sich auch der X^2-Wert als Maß gleicher Repräsentanz unter den Anfangsbuchstaben mit $X^2 = 104{,}5$ als weit über dem zulässigen theoretischen Wert von $\chi^2_9 = 16{,}9$ erweist. Eine besonders große Abweichung im prozentualen Anteil der aufgeführten Verben findet sich unter den Buchstaben B,I und D, die Wörterbücher stimmen überein bei S,T und E.

Bevor wir nun zum Vergleich von MAS und SSRLJa kommen, müssen wir anmerken, daß wir die Rubriken S und T als Repräsentanten für das Letztere nicht zufällig ausgewählt haben. Einmal hatten sich ja gerade für die Rubrik S verschiedene Verteilungen in MAS und WO ergeben; andererseits waren hier die relativen Anteile des lexikalischen Materials an den Wörterbüchern nahezu optimal. Der Grund für die Auswahl von I war zunächst, daß hier bei ausreichendem Stichprobenumfang die Wortzahl in Abhängigkeit von der Bedeutungszahl i besonders schnell absank, dann auch, daß gerade hier die größte Differenz im Wortumfang zwischen den Wörterbüchern vorlag.

An dieser Stelle muß bemerkt werden, daß für das vierbändige und das siebenbändige Akademiewörterbuch die Proportionen der Verbanzahlen zwischen den Rubriken I und S gut übereinstimmen (s. Tab. 2). Der experimentelle X^2-Wert ist hier mit $X^2 = 1{,}5$ unterhalb der kritischen Grenze $\chi^2_2 = 3{,}84$.

Charakteristisch für MAS und SSRLJa ist auch, daß nicht nur die absolute Anzahl ihrer Verbeinträge überhaupt, sondern auch die jeweilige Zahl der Wörter mit gleich vielen Lesarten nahe beieinander liegen (s. Tab. 1). Für kleine i sind die Verteilungen der Wörterbücher praktisch identisch; man überzeugt sich jedoch leicht davon, daß die Abnahme der mehrdeutigen Wörter im SSRLJa weniger steil verläuft als im MAS. Davon zeugt auch der Chi-Quadrat-Wert, der hier über der theoretischen Grenze liegt (für "I" ist $X^2 = 9{,}1 > \chi^2_3 = 7{,}8$; und für "S" finden wir $X^2 = 31{,}7 > \chi^2_9 = 16{,}9$). Zwar geben die Einträge im SSRLJa im Vergleich zum MAS und zum Lexikon von Ožegov einige Lesarten mehr an, aber es ist, wie zu Beginn des Abschnitts bemerkt, ohne theoretische Überlegungen zur Form der Verteilungsfunktion nicht möglich zu beurteilen, ob die Bedeutungszahlen in den Wörterbüchern MAS und WO zu niedrig angesetzt oder umgekehrt die im SSRLJa ungerechtfertigt hoch ausgefallen sind. Die weitere Analyse der experimentellen Resultate verschieben wir auf den Abschnitt 4 und kommen nun zu der theoretischen Fragestellung.

3. DAS GESETZ DES MAXIMALEN SEMANTISCHEN GEHALTS

Wir sondern nach einem beliebigen Merkmal aus der Menge des gesamten lexikalischen Materials eine Teilmenge mit relativ wenigen Elementen aus. Wir stellen nun die Hypothese auf, daß das grundlegende Konstruktionsprinzip der Lexik darin besteht, daß der semantische Gehalt des Subsystems sich durch die Gesamtzahl der Lesarten seiner Lexeme bestimmt und nicht davon abhängt, wie diese Lesarten auf die Elemente des Subsystems verteilt sind. Mit anderen Worten: wir behaupten, daß nicht das einzelne Wort, sondern jede seiner Bedeutungen durchschnittlich gleiche Information über die Welt der Denotate trägt.

Wir suchen nun die Wahrscheinlichkeit P_i dafür, daß sich das willkürlich abgegrenzte lexikalische Subsystem im i-ten Zustand befindet, d.h. daß es i Bedeutungen umfaßt. Wenn wir diese Wahrscheinlichkeitsverteilung kennen, können wir

1. die mittlere Anzahl von Subsystemen in einem gegebenen Zustand bestimmen (z.B. wenn wir als Subsystem das einzelne Wort wählen, können wir die Zahl der i-deutigen Wörter bestimmen),

2. die mittlere Bedeutungszahl eines Subsystems (z.B. die mittlere Anzahl von Lesarten pro Wort) angeben.

Um dieses Problem exakt zu lösen, müssen wir das Zeitverhalten der Wahrscheinlichkeitsfunktion betrachten. Die Menge der lexikalischen Einheiten jedoch, die in einem festen Zeitintervall zu einem lexikalischen Grundbestand hinzukommen oder fortfallen, wird im Vergleich zum Gesamtumfang des Wortschatzes relativ klein sein. Daher können wir annehmen, daß die Wahrscheinlichkeitsfunktion zu jedem Zeitpunkt nur wenig von dem synchronen Gleichgewicht abweicht, und uns bei der Betrachtung der Wahrscheinlichkeitsfunktion auf das Gleichgewicht beschränken, d.h. auf jenen Zustand, den sie annähme, wenn es uns gelänge, die äußeren Einflüsse auf sie aufzuheben bzw. das extralinguistische Moment "anzuhalten".

Um den korrekten Ausdruck für P_i zu finden, muß man vor allem beachten, daß jedes beliebige lexikalische Subsystem offen ist, und sein semantischer Gehalt sich sowohl linguistisch wie extralinguistisch bestimmt. Infolge ihrer Nichtabgeschlossenheit sind die einzelnen Subsysteme nicht vollständig unabhängig sondern lediglich quasiunabhängig.

Die Überlegung, inwieweit die Wechselwirkung zwischen dem Subsystem und seiner lexikalischen Umgebung (das ist die Menge aller lexikalischen Einheiten, die dem Subsystem nicht angehören) zu berücksichtigen ist, muß vor dem Hintergrund des dialektischen Gesetzes von der Einheit der Widersprüche angestellt werden. Einerseits ist unter Gleichgewichtsbedingungen die Summe der Bedeutungszahl des Subsystems und der seiner lexikalischen Umgebung (LU), I_u, annähernd konstant, also

$$i + I_u = I. \tag{3.1}$$

Insofern sind i und I_u eindeutig miteinander verknüpft (wenn das

Subsystem i Bedeutungen umfaßt, befindet sich seine LU in einem der möglichen Zustände mit der Bedeutungszahl $I_u = I - i$).

Andererseits ist $I_u \gg i$, so daß wir die Bedeutungszahl der LU als konstant ansehen können, wie groß i auch werden mag. Folglich ist der semantische Gehalt der LU unabhängig von dem des Subsystems; bei der Berechnung der Wahrscheinlichkeit P_i dafür, daß das Subsystem i und seine LU I_u Bedeutungen besitzen, multiplizieren wir daher die Einzelwahrscheinlichkeiten:

$$P_i = w(i)W(I_u) \tag{3.2}$$

Hier ist w(i) die Wahrscheinlichkeit für den i-ten Zustand des Subsystems und $W(I_u)$ die Wahrscheinlichkeit dafür, daß die LU gerade I_u Bedeutungen besitzt.

Nach unserem Grundpostulat ist der semantische Gehalt eines Subsystems eindeutig bestimmt. Es kann sich jedoch in verschiedenen Zuständen befinden, die sich durch die Verteilung der Bedeutungen auf die Elemente des Subsystems voneinander unterscheiden. Wir verwenden das Symbol $\Omega(i)$ für die Anzahl dieser Elementarzustände und bezeichnen sie als das statistische Gewicht des i-ten Zustands. Analog entspricht jedem I_u eine Gruppe von LU-Zuständen $\Omega_u(I_u)$. Alle Zustände, die ein und demselben semantischen Gehalt entsprechen, sind gleichwahrscheinlich, da dieser einzig von der Gesamtzahl an Lesarten abhängt. Mithin werden die Zustände mit größerem statistischen Gewicht auch wahrscheinlicher sein; genauer: die Wahrscheinlichkeit für einen Zustand ist proportional zu seinem statistischen Gewicht, $w(i) \sim \Omega(i)$, $W(I_u) \sim \Omega_u(I_u)$ und mit (3.1) und (3.2) erhalten wir für P_i

$$P_i = \Omega_u(I-i)\Omega(i). \tag{3.3}$$

Um nun der Wahrscheinlichkeitsverteilung (3.3) ihre konkrete Form zu geben, muß die Funktion $\Omega_u(I-i)$ gefunden werden. Dies wiederum ist wegen der geringen Größe von i gegenüber der gesamten Lexik (LU plus Subsystem) möglich. Man kann sich nämlich aus diesem Grund bei einer Potenzreihenentwicklung von $\Omega_u(I-i)$ auf das

erste Glied beschränken. Eine solche Potenzreihenentwicklung ist jedoch nicht unmittelbar möglich, weil die Anzahl der verschiedenen Zustände durch eine multiplikative und der semantische Gehalt durch eine additive Funktion ausgedrückt ist. (Die Anzahl der Zustände eines Systems unabhängiger Teile ist gleich dem Produkt der Zustandszahlen seiner Teile, während der semantische Gehalt die Summe der Bedeutungszahlen ist). Darum kann man Ω_u immer folgendermaßen schreiben:

$$\Omega_u(I-i) = e^{\sigma(I-i)} \tag{3.4}$$

Hier tritt eine Funktion $\sigma(I-i)$ auf, die wie der semantische Gehalt additiv ist.

Die Potenzreihenentwicklung von $\sigma(I-i)$ für kleines i unter Beschränkung auf das erste Glied ergibt:

$$\sigma(I-i) \approx \sigma(I) - \frac{\delta\sigma}{\delta I}i = \sigma - \lambda i \tag{3.5}$$

mit $\lambda = \frac{\delta\sigma}{\delta i}\Big|_{i=0}$. Damit finden wir für $\Omega_u(I-i)$:

$$\Omega_u(I-i) \approx e^{\sigma(I)}e^{-\lambda i} \tag{3.6}$$

und durch Einsetzen von (3.6) in (3.3)

$$P_i = Ce^{-\lambda i}\Omega(i). \tag{3.7}$$

Hier ist C das Produkt des Proportionalitätskoeffizienten und der Größe $e^{\sigma(I)}$, beide von i und den Eigenschaften des Subsystems unabhängig. Die Konstante erhält man aus folgender Normierungsbedingung:

$$\sum_{i=1}^{\infty} P_i = 1 \tag{3.8}$$

und somit ist

$$C = \frac{1}{\sum_{i=1}^{\infty} e^{-\lambda i}\Omega(i)} \tag{3.9}$$

Folglich erhält die Wahrscheinlichkeitsfunktion die endgültige Form

$$P_i = \frac{e^{-\lambda i}\Omega(i)}{\sum_{i=1}^{\infty} e^{-\lambda i}\Omega(i)} \tag{3.10}$$

Der Ausdruck (3.10) bestimmt die gesuchte Wahrscheinlichkeit P_i dafür, daß ein kleiner, quasiunabhängiger Teil des Gesamtwortbestands, ein lexikalisches Subsystem, einen semantischen Gehalt von i Lesarten hat. Die Verteilung (3.10) wurde im Jahr 1901 von Gibbs für mechanische Systeme gefunden; sie erhielt den Namen "Gibbs'sche" oder auch "Kanonische" Verteilung (s. z.B. Levič 1950). Daß unsere Überlegungen zur paradigmatischen Funktion zu eben dieser Gibbs'schen Verteilung führten, ist wegen ihrer Universalität nicht überraschend; deren Anwendbarkeitsbedingung ist nämlich bereits mit folgenden Forderungen erfüllt:

1. Die Existenz eines Makrosystems als Umgebung des betrachteten Systems;
2. Die Existenz einer vernachlässigbar schwachen Wechselwirkung zwischen dem Subsystem und seiner Umgebung.

Im Übrigen aber sind die Eigenschaften des Systems völlig beliebig.

Um zu unserer Fragestellung, der Suche nach der paradigmatischen Funktion, zurückzukommen, betrachten wir nun als Subsystem das einzelne Wort. Für ein solches - das einfachste - System ist das statistische Gewicht aller möglichen Zustände gleich 1, und der Ausdruck (3.10) vereinfacht sich auf

$$P_i = \frac{e^{-\lambda i}}{\sum_{i=1}^{\infty} e^{-\lambda i}} \tag{3.11}$$

Die Summe der geometrischen Reihe im Nenner von (3.11) mit $e^{-\lambda}$ als erstem Glied ist $\sum_{i=1}^{\infty} e^{-\lambda i} = \frac{e^{-\lambda}}{1-e^{-\lambda}}$. Das führt zur endgültigen paradigmatischen Wahrscheinlichkeitsfunktion:

$$P_i = (e^{\lambda} - 1)e^{-\lambda i} \quad , \tag{3.12}$$

Aus (3.12) folgt, daß die Wahrscheinlichkeit für ein i-deutiges Wort in einer Sprache lediglich von einem Parameter, den wir Verteilungsmodul nennen wollen, abhängt. Dieser Parameter muß positiv sein; denn nur mit $\lambda > 0$ konvergiert die Wahrscheinlichkeit P_i für $i \to \infty$ gegen Null.

Für die numerische Bestimmung des Verteilungsmoduls machen wir uns zunutze, daß die Entropie eines Systems mit seiner Annäherung an einen Gleichgewichtszustand wächst. Wie wir wissen, ist die Entropie eines Systems, das sich in einer abzählbaren Anzahl von Zuständen befinden kann, nach Shannon (1968)

$$H = -\sum_{i=1}^{n} P_i \ln P_i \quad , \tag{3.13}$$

Die Menge der möglichen Zustände unseres Subsystems ist nicht überabzählbar, und so bestimmen wir die Entropie durch die Reihe

$$H = -\sum_{i=1}^{\infty} (e^{\lambda}-1)e^{-\lambda i} \ln[(e^{\lambda}-1)e^{-\lambda i}]$$

$$= -(e^{\lambda}-1)[\ln(e^{\lambda}-1)\sum_{i=1}^{\infty} e^{-\lambda i} - \lambda\sum_{i=1}^{\infty} ie^{-\lambda i}] \quad , \tag{3.14}$$

Die erste Summe in (3.14) ist eine geometrische Reihe, wonach $\Sigma e^{-\lambda i} = \frac{1}{e^{\lambda}-1}$.

Um die zweite Summe zu berechnen, nutzen wir den Umstand, daß sie für beliebige $\lambda > 0$ eine Majorante hat; daher ist

$$\sum_{i=1}^{\infty} ie^{-\lambda i} = -\sum_{i=1}^{\infty} \frac{d}{d\lambda} e^{-\lambda i} = \frac{e^{\lambda}}{(e^{\lambda}-1)^2} \tag{3.15}$$

und die Entropie einer lexikalischen Einheit ergibt schließlich

$$H = \frac{\lambda e^{\lambda}}{e^{\lambda}-1} - \ln(e^{\lambda} - 1) \quad . \tag{3.16}$$

Unmittelbar ist eine Maximierung von (3.16) nach λ jedoch nicht möglich, da die Entropie unseres Subsystems für $\lambda \to 0$ unbegrenzt anwächst: $\lim H(\lambda) = \infty$. Das hängt damit zusammen, daß bei $\lambda \to 0$ die mittlere Bedeutungszahl der einzelnen lexikalischen Einheiten

$$M[i] = \sum_{i=1}^{\infty} ip_i = \frac{e^{\lambda}}{e^{\lambda}-1} \tag{3.17}$$

gegen unendlich strebt, was linguistisch natürlich uninterpretierbar ist. Um Divergenzen zu vermeiden, werden wir also nicht den absoluten Wert der Entropie maximieren, sondern die relative Entropie $h(\lambda)$; d.h. wir fordern, daß die mittlere Information einer Bedeutung des Subsystems ihr Maximum bei Annäherung an den Gleichgewichtszustand erreicht. Wir haben dann

$$h(\lambda) = \frac{H}{M[i]} = \lambda - \frac{(e^{\lambda}-1)\ln(e^{\lambda}-1)}{e^{\lambda}} \tag{3.18}$$

Durch Nullsetzen der ersten Ableitung von (3.18) erhalten wir die Bestimmungsgleichung für den Verteilungsmodul λ_0:

$$\ln(e^{\lambda_0} - 1) = 0, \tag{3.19}$$

deren einzige Lösung $\lambda_0 = \ln 2$ ist. Durch Einsetzen in (3.12) erhalten wir die Wahrscheinlichkeitsfunktion

$$P_i = 2^{-i} \quad . \tag{3.20}$$

Vor der Auswertung unserer Ergebnisse müssen wir noch besonders betonen, daß bei der Ableitung der Gleichung (3.20) keinerlei beschränkende Voraussetzungen über die lexikalischen Eigenschaften einer natürlichen Sprache gemacht wurden. Somit beschreibt diese Gleichung ein universelles Gesetz, gültig für beliebige Sprachen. Wir nennen es das Gesetz des maximalen semantischen Gehalts. Nach diesem Gesetz nimmt die *Anzahl der mehrdeutigen Wörter bei Annäherung des lexikalischen Systems an den Gleichgewichtszustand in Abhängigkeit von der Anzahl der Bedeutungen in geometrischer Progression mit dem Quotienten* $\frac{1}{2}$ *ab.* Das heißt: Alle mehrdeutigen Wörter insgesamt machen die Hälfte des Wortbestands der Sprache aus; die Menge der zweideutigen Wörter beträgt die Hälfte der der eindeutigen, die Menge der dreideutigen die Hälfte der der zweideutigen usw.

Aus dem Gesetz des maximalen semantischen Gehalts folgt ebenfalls unmittelbar, daß in einer beliebigen Sprache die mittlere Anzahl von Lesarten einer lexikalischen Einheit

$$M[i] = \frac{e^{\lambda_o}}{e^{\lambda_o} - 1} = 2 \qquad (3.21)$$

ist. Dieser Umstand charakterisiert den prinzipiellen Unterschied zwischen den formalen Sprachen, in denen jedes Zeichen genau eine Bedeutung besitzt, und den natürlichen Sprachen; er setzt offensichtlich auch die Grenze für die Möglichkeiten der automatischen Übersetzung. So bemerkte V.V. Nalimov (1974) zutreffend: "Die Mehrdeutigkeit der Sprache kommt die Menschen teuer zu stehen".

Interessant ist noch, daß bei $\lambda_o = \ln 2$ die mittlere Entropie pro Lesart ebenfalls ln 2 ist, d.h. gleich der Entropie des einfachsten Systems, das sich nur in einem von zwei gleichmöglichen Zuständen befinden kann, d.h. dessen Elemente mit gleicher Wahrscheinlichkeit die Werte Null und Eins annehmen.

4. VERGLEICH DER THEORETISCHEN WAHRSCHEINLICHKEITSFUNKTION MIT DEN EMPIRISCHEN VERTEILUNGEN

Die theoretische Wahrscheinlichkeitsfunktion (3.20) aus Abschnitt 3 charakterisiert die Lexik der natürlichen Sprache insgesamt und nicht getrennt nach Anfangsbuchstaben. Daher muß der Vergleich mit solchen empirischen Verteilungen durchgeführt werden, die auf für das jeweilige Wörterbuch insgesamt repräsentativen Stichproben beruhen. Demgemäß wurde die Stichprobe nach den Anfangsbuchstaben A bis I, S und T für das WO (Tab. 1) durch Stichproben nach allen verbleibenden Anfangsbuchstaben ergänzt. Die Erhebung erfolgte wie anfangs beschrieben. In Tab. 3 ist die resultierende Stichprobe dargestellt.

Tabelle 3. Empirische und theoretische Verteilungen mehrdeutiger Wörter

	1	2	3	4	5	6	7	8	9	10	11
n_iWO"A-Ja"	5845	2420	675	289	123	64	30	14	16	15	11
n_iWO"IKS"	822	333	93	42	16	6	5	4	0	5	3
n_iSSRLJa	1421	672	238	122	79	50	32	23	22	12	40
ν_iWO"A-Ja"	0,615	0,254	0,071	0,030	0,013	0,007	0,003	0,002	0,002	0,002	0,001
ν_iWO"IKS"	0,614	0,249	0,070	0,031	0,012	0,005	0,004	0,003	0,000	0,004	0,003
ν_iSSRLJa	0,525	0,248	0,084	0,045	0,029	0,018	0,012	0,009	0,008	0,004	0,023
P_i theor.	0,500	0,250	0,125	0,063	0,031	0,016	0,008	0,004	0,002	0,001	0,002

Allerdings ergab die Analyse, daß die Stichprobe mit I, K und S schon für den gesamten Bestand des Wörterbuchs repräsentativ war. In Tab. 3 wurden in den Zeilen ν_i WO "A - Ja" und ν_i WO "I,K,S" die entsprechenden beobachteten Häufigkeiten sowohl für den Gesamtbestand des Wörterbuchs wie auch für die Buchstaben I, K und S allein aufgeführt.

Man überzeugt sich leicht, daß die Häufigkeiten sehr nahe aneinander liegen, und ihr Unterschied als zufällig angesehen werden kann. Die Überprüfung der Hypothese, daß die Auswahl I, K, S

für das gesamte Wörterbuch repräsentativ ist, ergab in der Tat $\chi^2 = 4{,}78 < \chi^2_7 = 14{,}1$ für $r = s - 1 = 7$ Freiheitsgrade (vgl. Abschnitt 2). Für das SSRLJa wurde die Stichprobe nach I und S lediglich durch eine Stichprobe nach K erweitert. Die sich daraus ergebende Verteilung ist ebenfalls in Tab. 3 dargestellt. Schließlich sind in der Zeile $P_{i(th)}$ die theoretischen Wahrscheinlichkeitswerte für das Antreffen von i-deutigen Wörtern, berechnet nach (3.20), angegeben.

Unsere Daten und die dazugehörige theoretische Kurve sind auch in Abb. 1 wiedergegeben. Auf der Abzisse ist die Bedeutungszahl i abgetragen, und auf der Ordinate $\mathrm{ld}\ P_i$. Die Skalen wurden so gewählt, weil auf diese Weise die theoretische Abhängigkeit der Variablen graphisch als Gerade erscheint, die auf der Winkelhalbierenden des II. und IV. Quadranten liegt. Durch Logarithmieren zur Basis 2 erhalten wir nämlich aus (3.20)

$$\mathrm{ld}\ P_i = - i \qquad (4.1)$$

Wie aus Abb. 1 ersichtlich ist, sind die Punkte für die Daten aus WO und SSRLJa in unmittelbarer Nähe der theoretischen Geraden verteilt. Die Übereinstimmung der theoretischen Ableitung mit den Daten muß als überzeugend angesehen werden, besonders wenn man bedenkt, daß die theoretische Verteilung keinen aus den Daten geschätzten Parameter enthält, d.h. vollkommen unabhängig von den empirischen Daten aufgestellt wurde, was nur äußerst selten gelingt.

Die Abweichungen der Punkte von der theoretischen Geraden verhalten sich für die beiden Wörterbücher in gewissem Maße gleich. So ist die Anzahl der eindeutigen Wörter in beiden Fällen etwas erhöht. Die beobachteten Häufigkeiten der Wörter mit relativ wenigen Bedeutungen ($2 \leq i < 7$) sind geringer als theoretisch erwartet. Im weiteren Verlauf nähern sie sich mit anwachsendem i wiederum an die Gerade an und ergeben schließlich bei großem i für beide Wörterbücher eine systematische Abweichung zugunsten erhöhter Häufigkeiten.

Daß die beobachteten und die theoretisch erwarteten Häufigkeiten in ihrem Abweichungsverhalten so gleichartig sind, läßt

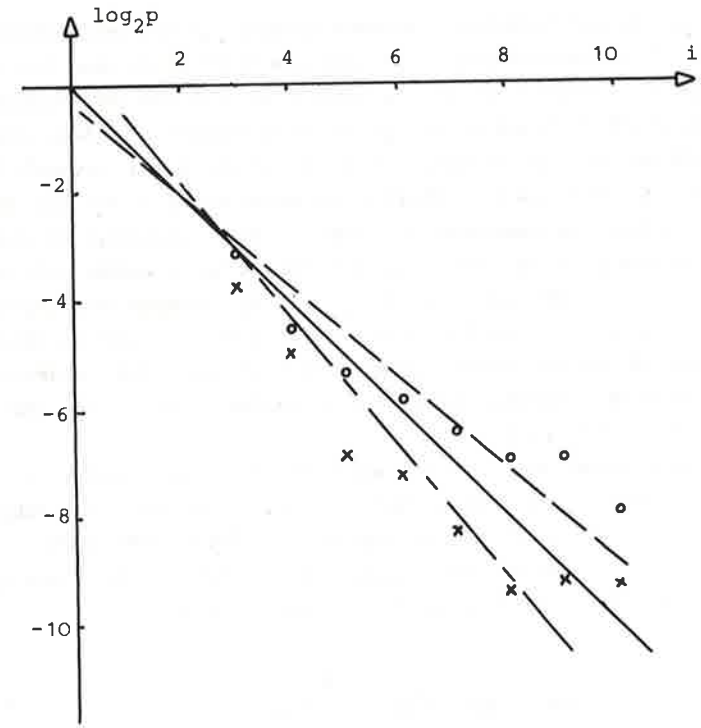


Abb. 1. Theoretische und empirische Verteilung mehrdeutiger Wörter in Abhängigkeit von der Anzahl der Bedeutungen
(o für SSRLJA, x für WO)

uns für beide Wörterbücher die gleichen Ursachen vermuten. Wesentlich sind hier die systematischen Fehler, die bei der Bestimmung der Bedeutungszahl von Wörtern mit stark entwickeltem Polymorphismus auftreten. Sie sind dadurch bedingt, daß die Probleme bei der Definition des Wortbegriffs im grammatischen wie semantischen Sinn sehr komplex sind. Andererseits ist nicht ausgeschlossen, daß die Abweichung zugunsten erhöhter Häufigkeiten für eindeutige Wörter in gewissem Maße damit zusammenhängt, daß die wahre paradigmatische Verteilung selbst vom Gleichgewicht abweicht.

Bezüglich der konkreten Gegenüberstellung der Wörterbücher mit einem "Idealwörterbuch", in dem die Wortverteilung den theoretischen Forderungen genügt, zeigen wir, daß die empirischen Punkte beträchtlich näher zu den entsprechenden Geraden, die als unterbrochene Linien in Abb. 1 eingezeichnet sind, liegen. Diese Geraden wurden aus der Überlegung aufgestellt, daß der prozentuale Gehalt an mehrdeutigen Wörtern eine Exponentialfunktion gemäß Gleichung (3.12) ist. Für den Parameter λ wurde allerdings nicht der reintheoretische Wert $\lambda_o = \ln 2$, sondern die empirische Größe $\tilde{\lambda}$ eingesetzt, die für jedes Wörterbuch aus der geschilderten Untersuchung errechnet wurde. Die Schätzung des Parameters $\tilde{\lambda}$ wurde nach der Maximum-Likelihood-Methode (MLM) (vgl. van der Waerden 1960) durchgeführt.

Wir bezeichnen mit P_{is} die Wahrscheinlichkeit dafür, daß das s-te Wort des Wörterbuchs gerade i Bedeutungen hat. Wir erhalten durch Multiplikation der Einzelwahrscheinlichkeiten dafür, daß das erste Wort i_1 Bedeutungen hat, das zweite i_2, das s-te i_s usw., schließlich das n-te Wort i_n, das Produkt

$$P = P_{i1}P_{i2}\cdots P_{is}\cdots P_{in} = \prod_{s=1}^{n} P_{is} \tag{4.2}$$

Setzen wir in (4.2) den Ausdruck (3.12) ein, so erhalten wir

$$P = \prod_{s=1}^{n} (e^{\lambda}-1)e^{-\lambda i_s} \tag{4.3}$$

Durch Logarithmieren von (4.3) ergibt sich

$$L = \sum_{s=1}^{n} \ln(e^{\lambda}-1)e^{-\lambda i_s} = \sum_{s=1}^{n} [-\lambda i_s + \ln(e^{\lambda}-1)]. \tag{4.4}$$

Nach dem MLM-Schätzverfahren fordern wir

$$\left.\frac{dL}{d\lambda}\right|_{\lambda=\tilde{\lambda}} = 0 . \tag{4.5}$$

So erhalten wir zur Berechnung von $\tilde{\lambda}$ die Gleichung

$$\sum_{s=1}^{n} i_s + \sum_{s=1}^{n} \frac{e^{\tilde{\lambda}}}{e^{\tilde{\lambda}}-1} = 0 \tag{4.6}$$

oder gleichbedeutend

$$\tilde{\lambda} = \ln \frac{\bar{i}}{\bar{i} - 1} \tag{4.7}$$

wobei $\bar{i}$ das arithmetische Mittel der beobachteten Zufallsvariablen (der Bedeutungszahl eines einzelnen Wortes) bezeichnet:

$$\bar{i} = \frac{1}{n} \sum_{s=1}^{n} i_s . \tag{4.8}$$

Durch Einsetzen von (4.7) in (3.12) und Logarithmieren zur Basis 2 erhalten wir die Gleichungen der Geraden in Abb. 1:

$$\mathrm{ld}\, P_i = \mathrm{ld}(\bar{i} - 1)^{-1} - \mathrm{ld} \frac{\bar{i}}{\bar{i} - 1} \tag{4.9}$$

Die Berechnung von $\bar{i}$ aus den Daten von Tab. 3 ergab für das WO $\bar{i}_1 = 1{,}69$ und für das SSRLJa $\bar{i}_2 = 2{,}25$ und damit die empirischen Schätzungen $\tilde{\lambda}_1 = \ln 2{,}45$ bzw. $\tilde{\lambda}_2 = \ln 1{,}8$. Diese Resultate besagen nun:

1. Für beide Wörterbücher sind die empirischen Werte des Verteilungsmoduls, wie zu erwarten, nahe am Gleichgewichtswert des Parameters, nämlich $\lambda_o = \ln 2$;
2. Die empirische Verteilung von SSRLJa ist näher an der theoretischen (mittlere Abweichung $\frac{\Delta i}{\bar{i}} = \frac{i-2}{\bar{i}} \approx 10\%$) als die vom WO ($\frac{\Delta i}{\bar{i}} \approx 20\%$). Während das WO im Mittel eine eigentümliche Tendenz zur Abweichung zugunsten geringerer Bedeutungszahlen hat, ist für SSRLJa umgekehrt eine gewisse Erhöhung im Vergleich zu den theoretischen Zahlen charakteristisch.

5. DAS HOMONYMIEPROBLEM UND DAS GESETZ DES MAXIMALEN SEMANTISCHEN GEHALTS

Die vorgelegten Resultate haben nicht nur theoretische Bedeutung, sondern sie können auch bei der Lösung einer ganzen Reihe alltäglicher linguistischer Probleme praktische Anwendung finden. Ein solches Problem ist die Bestimmung der Homonyme im Wörterbuch. Wie schon häufig in der linguistischen Literatur beklagt wurde, leiden die heute bekannten semantischen und nichtsemantischen Kriterien für die Wortidentität an ein und demselben Unvermögen: die Autoren beschreiben sie anhand einiger, glücklich gewählter Beispiele, und der Lexikograph, der sie bei seiner Arbeit zu verwenden versucht, stößt auf eine so große Zahl von Ausnahmen, daß der Wert des Kriteriums in Frage gestellt ist.

Die Forderung nach interner Widerspruchsfreiheit legt der paradigmatischen Funktion der Homonyme eine starke Beschränkung auf. Wir verstehen unter Homonymen zwei verschiedene Wörter mit identischer materieller Form; das einzige Merkmal, das sie von allen anderen Wörtern unterscheidet, ist also die Gemeinsamkeit ihrer Form. Daher müssen zwei Homonyme - als semantisch unabhängige Einheiten - der gleichen paradigmatischen Funktion unterliegen wie ein Subsystem aus zwei zufällig gewählten Wörtern, das wir von nun ab binär nennen wollen.

Das einfachste binäre Subsystem enthält zwei Bedeutungen. Die Bildung eines solchen Subsystems ist eigentlich die Überschneidung der beiden Ereignisse A_1 (das erste Wort des Subsystems ist eindeutig) und A_2 (das zweite Wort ist eindeutig). Da nun die Bedeutungszahl des einen nicht von der Bedeutungszahl des anderen Wortes abhängt, und da $P(A_1) = P(A_2) = P_1$ (d.h. die Wahrscheinlichkeit dafür, in der Sprache ein eindeutiges Wort anzutreffen), muß nach dem Satz von der Multiplikativität der Wahrscheinlichkeiten die Wahrscheinlichkeit für das Antreffen eines einfachen binären Subsystems $\tilde{P}_2 = P_1^2$ sein. Es ist nicht schwer, analog mit Hilfe des Additivitäts- und des Multiplikativitätssatzes den Ausdruck für die Wahrscheinlichkeit aufzustellen, daß das binäre Subsystem genau i Bedeutungen umfaßt:

$$\tilde{P}_i = \sum_{s=1}^{i} P_s P_{i-s}. \qquad (5.1)$$

Infolge der Größe der Stichproben dürften die Häufigkeiten ν_i von Vorkommen i-deutiger Wörter (wenigstens im Gebiet kleiner i) nicht sehr von den entsprechenden Wahrscheinlichkeiten differieren, so daß die Beziehung (5.1) mit hinreichender Genauigkeit auch für die Häufigkeiten benutzbar sein muß. Dieser Beziehung müssen auch die Häufigkeiten f_i von Homonymen mit der Gesamtbedeutungszahl i genügen, wie sie für das WO (ganz) und MAS (Stichproben A bis Z, I,S,T) in Tab. 4 angeführt sind. Angegeben sind dort die absolute Anzahl der in diesen Wörterbüchern isolierten Homonyme m_i und die erwarteten Häufigkeiten $\tilde{f}_i$, wie sie aus den empirischen Verteilungen der eindeutigen Wörter berechnet wurden.

Tabelle 4. Verteilung homonymer Verben in Abhängigkeit von der Zahl ihrer Bedeutungen

		2	3	4	5	6	i ≥ 7	n
WO	m_i	151	85	44	18	10	9	317
	f_i	0,476	0,268	0,139	0,057	0,031	0,028	
	$\tilde{f}_i$	0,381	0,314	0,151	0,073	0,037	0,045	
MAS	m_i	34	21	6	11			72
	f_i	0,458	0,292	0,083	0,159			
	$\tilde{f}_i$	0,375	0,289	0,145	0,182			

Die Untersuchung zeigte, daß für MAS bei relativ kleiner Gesamtzahl an isolierten Homonymen (n = 72) der Unterschied zwischen empirischen und theoretisch erwarteten Häufigkeiten als zulässig angesehen werden kann: $X^2 = 4{,}52 < 7{,}82 = \chi^2_3$. Im vorliegenden Fall wurde zur Berechnung des experimentellen X^2-Wertes anstelle der Formel (2.1) der gleichwertige Ausdruck

$$X^2 = \sum_{i=1}^{s} \frac{(m_i - n\tilde{f}_i)^2}{n\tilde{f}_i} \qquad (5.2)$$

verwendet (vgl. Dunin-Barkovskij/Smirnov 1955). Fast der gleiche

Unterschied in der Häufigkeit zeigte sich auch für das WO. Allerdings kann man ihn bei n = 317 nicht mehr als zufällig ansehen; in diesem Fall ist $X^2 = 17{,}04 > 11{,}07 = \chi^2_5$, und damit liegt in der Struktur des WO eine Widersprüchlichkeit vor.

In seiner ganzen Komplexität präsentiert sich das Homonymieproblem bei der semantischen Betrachtung der Wortidentitätsfrage. In der lexikographischen Praxis ergibt sich bei der Untersuchung der Polysem - /Homonym - Abgrenzung, geht man von bestimmten Kriterien aus, eine Gruppe von lexikalischen Einheiten, deren semantische Wortidentität vom Standpunkt der Wörterbuchautoren aus verletzt ist. Die zeitgenössische Linguistik bietet allerdings, wie oben bemerkt, keine eindeutigen Abgrenzungskriterien an, was bei verschiedenen Definitionswörterbüchern zu divergierenden Entscheidungen führt.

Die vorgestellten Ergebnisse gestatten es, ein objektives Testkriterium dafür vorzuschlagen, wann die Polysem-Abtrennung als beendet angesehen werden muß. Um Mißverständnisse zu vermeiden muß jedoch betont werden, daß das angesprochene Kriterium notwendig aber nicht hinreichend ist und sich dabei auf die Gesamtheit aller untersuchten lexikalischen Einheiten bezieht.

Um Polyseme und Homonyme voneinander abzugrenzen, kann man sich den Unterschied zwischen den Verteilungsfunktionen der Bedeutungszahlen von gewöhnlichen eindeutigen Wörtern und Homonymen zunutze machen. Für die Homonyme erhält man die Verteilungsfunktion aus dem allgemeinen Ausdruck (3.10) und dem statistischen Gewicht $\Omega(i) = i - 1$ für das gegebene Subsystem. Nehmen wir wieder $\lambda = \lambda_o = \ln 2$, so erhalten wir

$$\tilde{P}_i = (i - 1)2^{-i} \qquad (5.3)$$

Für das eindeutige Wort als Subsystem wird die Wahrscheinlichkeitsfunktion durch die Gleichung

$$P_i = \frac{2^{-i}}{1 - P_1} = 2 \cdot 2^{-i} \qquad (5.4)$$

beschrieben, die sich von (3.20) dadurch unterscheidet, daß hier

nur lexikalische Einheiten mit einer Bedeutungszahl von $i \geq 2$ in Betracht gezogen werden. Die theoretischen Kurven der Gleichungen (5.3) und (5.4) zeigt Abb. 2.

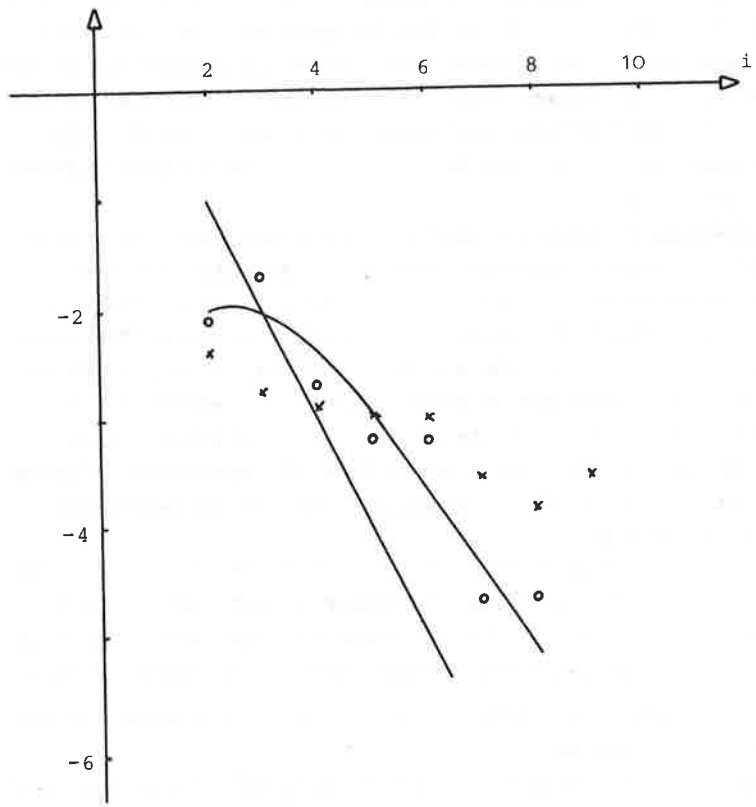


Abb. 2. 1-o- die theoretische Kurve und die experimentellen Punkte für Homonyme

2-x- die theoretische Gerade und die experimentellen Punkte für Verben in statu nascendi

Nehmen wir nun an, uns liegt eine Gruppe von Einheiten vor, deren Homonymhaltigkeit wir untersuchen wollen. Wenn wir die empirische Verteilung bestimmt haben, können wir sie mit der theoretischen Verteilung (5.4) vergleichen. Falls ein Unterschied zwischen empirischer und theoretischer Verteilung auftritt, sind wir genötigt, Wortidentität für die überprüften Einheiten (innerhalb der Gesamtheit) anzunehmen, d.h. das Trennungsmerkmal reicht nicht aus, um einen semantischen Bruch zu bestätigen. Fallen aber empirische und theoretische Verteilung zusammen, kann das als nachträglicher Beweis für die Richtigkeit der getesteten Hypothese angesehen werden.

Als praktisches Beispiel werden wir das vorgeschlagene Kriterium auf die Verben anwenden, die im Homonymielexikon von Achmanova (1974) in die Gruppe III (Polyseme) eingeordnet sind. Die Verfasserin führt in diesem Lexikon neben sicheren Homonymen auch solche an, die sich, wie sie es ausdrückt, in statu nascendie befinden. Die empirischen Verteilungen sind sowohl für die Verben, die von der Autorin als Homonyme (H) aufgefaßt werden, wie auch für die Verben, die sich im Aufspaltungsprozeß befinden (H^x), in Tab. 5 und in Abb. 2 angegeben (die Bedeutungszahlen basieren auf dem SSRLJa).

Wie man sieht, gruppieren sich die Punkte für die Verben mit dem Vermerk "H" dicht um die Homonymiekurve, was auch der Chi-Quadrat-Wert $X^2 = 3{,}43 < 11{,}1 = \chi^2_5$ bestätigt, der aus (5.2) nach Austausch von $\tilde{f}_i$ mit dem theoretischen Wert $\tilde{P}_i$ berechnet wurde. Dieses Ergebnis kann als weitere, unabhängige Bestätigung für die vorgelegte Theorie gelten.

Die Punkte für die Verben mit dem Vermerk "H^x" gruppieren sich um eine Gerade, deren Lage sich wesentlich von der theoretischen Geraden für die mehrdeutigen Verben und auch von der Homonymiekurve unterscheidet. Letzteres bestätigt Achmanovas Behauptung, daß die Beziehungen zwischen den Bedeutungen der von ihr isolierten Verben noch nicht für die gewöhnliche Mehrdeutigkeit charakteristisch sind.

Tabelle 5. Empirische Verteilung der Zahl der Homonyme und der Wörter, die ihre semantische Identität allmählich verlieren

i		2	3	4	5	6	7	n			
WO	m_i	6	8	4	3	3	5	29			
	f_i	0,207	0,276	0,138	0,103	0,103	0,172				
	$\tilde{P}_i$	0,250	0,250	0,188	0,125	0,078	0,109				
i		2	3	4	5	6	7	8	9	10	n
WO	m_i	14	11	10	9	9	6	5	6	10	80
	f_i	0,175	0,138	0,125	0,112	0,112	0,075	0,063	0,075	0,125	
	$\tilde{P}_i$	0,250	0,250	0,188	0,125	0,078	0,047	0,027	0,016	0,019	

Die in der vorliegenden Arbeit behandelten Fragen hängen auch unmittelbar mit dem Problem der Wortidentität auf grammatischer Ebene zusammen. Selbstverständlich werden sich die Resultate etwas unterscheiden, wenn man bei der Bestimmung der Bedeutungszahlen der einzelnen Lexeme von verschiedenen Vorstellungen über die regulären grammatischen Spielarten der Worte ausgeht, also darüber, welcher Grad von "Grammatikalität" und Regularität entscheidet, ob es sich bei einer Spielart um eine "grammatische Form" oder um ein weiteres Wort handelt. Die Analyse ergab jedoch, daß sich die empirischen Verteilungen, die auf verschiedenen Methoden zur Bestimmung der Bedeutungszahlen beruhen, wenig voneinander unterscheiden und sehr nahe an der theoretischen Verteilung nach Gleichung (3.20) liegen, sich also in Übereinstimmung mit dem Gesetz des maximalen semantischen Gehalts befinden. Außerdem kann die Nähe empirischer Verteilungen zu der theoretischen

Verteilung offensichtlich als Kriterium für die Richtigkeit der einen oder anderen Auffassung dienen - wie bei strittigen Fragen zur Wortidentität, wie in unserem Fall, oder z.B. bei der Frage, ob man die verschiedenen Handlungsarten / genera verbi als selbständige Wörter oder als Wortformen zählen sollte.

LITERATUR

ACHMANOVA, O.S., Slovar' omonimov russkogo jazyka Moskva, Sovetskaja enciklopedija 1974

ANDREEV, N.D., Statističeskie i kombinatornye metody v teoretičeskom i prikladnom jazykoznanii. Leningrad, Nauka 1967

ARAPOV, M.V., CHERC, M.M., Matematičeskie metody v istori - českoj lingvistike. Moskva, Nauka 1974

ARAPOV, M.V., EFIMOVA, E.N., ŠREJDER, Ju.A., O smysle rangovych raspredelenij. Naucno-techničeskaja informacija 2, Nr. 1, 1975

BONDARKO, A.V., Russkij glagol. Leningrad 1967

DUNIN-BARKOVSKIJ, I.V., SMIRNOV, P.V., Teorija verojatnostej i matematičeskaja statistika v technike. Moskva, GITTL 1955

LEVIČ, V.G., Vvedenie v statističeskuju fiziku. Leningrad, GITTL 1950

NALIMOV, V.V., Verojatnostnaja model' jazyka. Moskva, Nauka 1974

PIOTROWSKI, R.G., Informacionnye izmerenija jazyka. Moskva, Nauka 1968

ŠENNON, K., Raboty po teorii informacii i kibernetike. Moskva, IL 1963

ZIPF, G.K. Human behavior and the principle of least effort. Cambridge, Mass., Addison-Wesley 1949

QUANTITATIVE LINGUISTICS

Appeared

Vol. 1. Altmann, G. (Ed.), Glottometrika 1
Vol. 2. Grotjahn, R., Linguistische uns statistische Methoden in Metrik und Textwissenschaft
Vol. 3. Grotjahn, R. (Ed.), Glottometrika 2
Vol. 4. Strauß, U., Struktur und Leistung der Vokalsysteme
Vol. 5. Matthäus, W. (Ed.), Glottometrika 3
Vol. 6. Grotjahn, R., Hopkins, E. (Eds.), Empirical Research on Language Teaching and Language Acquisition
Vol. 7. Altmann, G., Lehfeldt, W., Einführung in die quantitative Phonologie 1
Vol. 8. Altmann, G., Statistik für Linguisten 1
Vol. 9. Hopkins, E., Grotjahn, R. (Eds.), Studies in Language Teaching and Language Acquisition
Vol. 10. Skorochod'ko, E.F., Semantische Relationen im Lexikon und in Texten
Vol. 11. Grotjahn, R. (Ed.), Hexameter Studies
Vol. 12. Rieger, B. (Ed.), Empirical Semantics. A Collection of New Approaches in the Field, Vol. 1
Vol. 13. Rieger, B. (Ed.), Empirical Semantics. A Collection of New Approaches in the Field, Vol. 2
Vol. 14. Lehfeldt, W., Strauß, U. (Eds.), Glottometrika 4
Vol. 15. Orlov, Ju.K., Boroda, M.G., Nadarejsvili, I.S., Sprache, Text, Kunst. Quantitative Analysen
Vol. 16. Guiter, H., Arapov, M.V. (Eds.), Studies on Zipf's Law

In Preparation

Alekseev, P., Statistische Lexikographie
Altmann, G., Diskrete Wahrscheinlichkeitsverteilungen
Arapov, M.V., Cherc, M.M., Mathematische Methoden in der historischen Linguistik
Boy, J., Mathematische Grundverfahren für Linguistik
Boy, J., Phonetische Dechiffrierung von Buchstaben
Brainerd, B. (Ed.), Historical Linguistics
Frank, H. (Hrsg.), Interlinguistik
Gindin, S. (Ed.), Dialectology
Lesochin, M.M., Piotrowski, R.G., Mathematical Models in Quantitative Linguistics
Matthäus, W. (Ed.), Psycholinguistics
Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A., Mathematische Linguistik

The series publishes
- Mixed volumes
- Monothematic volumes
- Textbooks
- Monographs
- Frequency dictionaries

Contributions can be sent to G. Altmann, Sprachwissenschaftliches Institut der RUB, Postfach 102148, 4630 Bochum, West Germany
Orders for individual volumes or the entire series should be directed to Studienverlag Dr. N. Brockmeyer, Querenburger Höhe 281, 4630 Bochum, West Germany.