Peter Grzybek (ed.): *Contributions to the Science of Text and Language.*
Dordrecht: Springer, 2007, pp. 15–90

HISTORY AND METHODOLOGY OF WORD LENGTH STUDIES

The State of the Art

Peter Grzybek

1. Historical roots

The study of word length has an almost 150-year long history: it was on August 18, 1851, when Augustus de Morgan, the well-known English mathematician and logician (1806–1871), in a letter to a friend of his, brought forth the idea of studying word length as an indicator of individual style, and as a possible factor in determining authorship. Specifically, de Morgan concentrated on the number of letters per word and suspected that the average length of words in different Epistles by St. Paul might shed some light on the question of authorship; generalizing his ideas, he assumed that the average word lengths in two texts, written by one and the same author, though on different subjects, should be more similar to each other than in two texts written by two different individuals on one and the same subject (cf. Lord 1958).

Some decades later, Thomas Corwin Mendenhall (1841–1924), an American physicist and metereologist, provided the first empirical evidence in favor of de Morgan's assumptions. In two subsequent studies, Mendenhall (1887, 1901) elaborated on de Morgan's ideas, suggesting that in addition to analyses "based simply on mean word-length" (1887: 239), one should attempt to graphically exhibit the peculiarities of style in composition: in order to arrive at such graphics, Mendenhall counted the frequency with which words of a given length occur in 1000-word samples from different authors, among them Francis Bacon, Charles Dickens, William M. Thackerey, and John Stuart Mill. Mendenhall's (1887: 241) ultimate aim was the description of the "normal curve of the writer", as he called it:

> [...] it is proposed to analyze a composition by forming what may be called a 'word spectrum' or 'characteristic curve', which shall be a graphic representation of the arrangement of words according to their length and to the relative frequency of their occurrence.

Figure 2.1, taken from Mendenhall (1887: 237), illustrates, by way of an example, Mendenhall's achievements, showing the result of two 1000-word samples from Dickens' *Oliver Twist*: quite convincingly, the two curves converge to an astonishing degree.

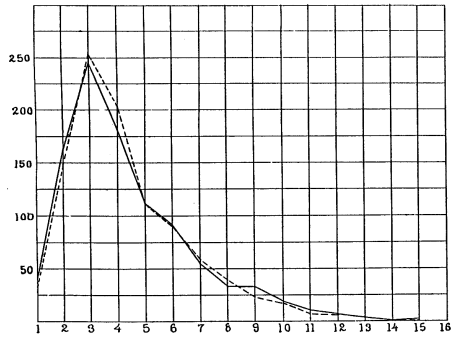


Figure 2.1: Word Length Frequencies in Dickens' *Oliver Twist* (Mendenhall 1887)

Mendenhall (1887: 244) clearly saw the possibility of further applications of his approach:

> It is hardly necessary to say that the method is not necessarily confined to the analysis of a composition by means of its mean word-length: it may equally well be applied to the study of syllables, of words in sentences, and in various other ways.

Still, Mendenhall concentrated solely on word length, as he did in his follow-up study of 1901, when he continued his earlier line of research, extending it also to include selected passages from French, German, Italian, Latin, and Spanish texts.

As compared to the mere study of mean length, Mendenhall's work meant an enormous step forward in the study of word length, since we know that a given mean may be achieved on the basis of quite different frequency distributions. In fact, what Mendenhall basically did, was what would nowadays rather be called a frequency analysis, or frequency distribution analysis. It should be mentioned, therefore, that the mathematics of the comparison of frequency distributions was very little understood in Mendenhall's time. He personally was mainly attracted to the frequency distribution technique by its resemblance to spectroscopic analysis.

Figure 2.2, taken from Mendenhall (1901: 104) illustrates the curves from two passages by Bacon and Shakespeare. Quite characteristically, Mendenhall's conclusion was a suggestion to the reader: "The reader is at liberty to draw any conclusions he pleases from this diagram."

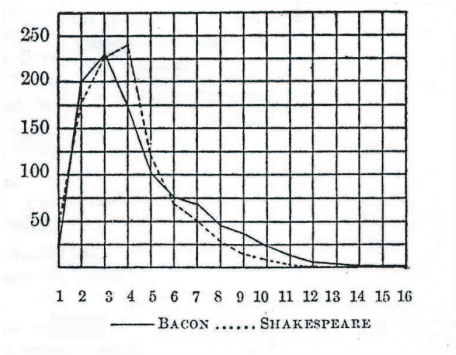


Figure 2.2: Word Length Frequencies in Bacon's and Shakespeare's Texts (Mendenhall 1901)

On the one hand, one may attribute this statement to the author's 'scientific caution', as Williams (1967: 89) put it, discussing Mendenhall's work. On the other hand, the desire for calculation of error or significance becomes obvious, techniques not yet well developed in Mendenhall's time.

Finally, there is another methodological flaw in Mendenhall's work, which has been pointed out by Williams (1976). Particularly as to the question of authorship, Williams (1976: 208) emphasized that before discussing the possible significance of the Shakespeare–Bacon and the Shakespeare–Marlowe controversies, it is important to ask whether any differences, other than authorship, were involved in the calculations. In fact, Williams correctly noted that the texts written by Shakespeare and Marlowe (which Mendenhall found to be very similar) were primarily written in blank verse, while all Bacon's works were in prose (and were clearly different). By way of additionally analyzing works by Sir Philip Sidney (1554–1586), a poet of the Elizabethan Age, Williams (1976: 211) arrived at an important conclusion:

> There is no doubt, as far as the criterion of word-length distribution is concerned, that Sidney's prose more closely resembles prose of Bacon than it does his own verse, and that Sidney's verse more closely resembles the verse plays of Shakespeare than it does his own prose. On the other hand, the pattern of difference between Shakespeare's verse and Bacon's prose is almost exactly comparable with the difference between Sidney's prose and his own verse.

Williams, too, did not submit his observations to statistical testing; yet, he made one point very clear: word length need not, or not only, or perhaps not even primarily, be characteristic of an individual author's style; rather word length, and word length frequencies, may be dependent on a number of other factors, genre being one of them (cf. Grzybek et al. 2005, Kelih et al. 2005).

Coming back to Mendenhall, his approach should thus, from a contemporary point of view, be submitted to cautious criticism in various aspects:

(a) *Word length is defined by the number of letters per word.*– Still today, many contemporary approaches (mainly in the domain of computer sciences), measure word length in the number of letters per word, not paying due attention to the arbitrariness of writing systems. Thus, the least one would expect would be to count the number of sounds, or phonemes, per word; as a matter of fact, it would seem much more reasonable to measure word length in more immediate constituents of the word, such as syllables, or morphemes. Yet, even today, there are no reliable systematic studies on the influence of the measuring unit chosen, nor on possible interrelations between them (and if they exist, they are likely to be extremely language-specific).

(b) *The frequency distribution of word length is studied on the basis of arbitrarily chosen samples of 1000 words.*– This procedure, too, is often applied, still today. More often than not, the reason for this procedure is based on the statistical assumption that, from a well-defined sample, one can, with an equally well-defined degree of probability, make reliable inferences about some totality, usually termed population. Yet, as has been repeatedly shown, studies along this line do not pay attention to a text's homogeneity (and consequently, to data homogeneity). Now, for some linguistic questions, samples of 1000 words may be homogeneous – for example, this seems to be the case with letter frequencies (cf. Grzybek/Kelih/Altmann 2004). For other questions, particularly those concerning word length, this does not seem to be the case – here, any selection of text segments, as well as any combination of different texts, turns out to be a "quasi text" destroying the internal rules of textual self-regulation. The very same, of course, has to be said about corpus analyses, since a corpus, from this point of view, is nothing but a quasi text.

(c) *Analyses and interpretations are made on a merely graphical basis.*– As has been said above, the most important drawback of this method is the lack of objectivity: no procedure is provided to compare two frequency distributions, be it the comparison of two empirical distributions, or the comparison of an empirical distribution to a theoretical one.

(d) *Similarities (homogeneities) and differences (heterogeneities) are unidimensionally interpreted.*– In the case of intralingual studies, word length frequency distributions are interpreted in terms of authorship, and in the case of interlingual comparisons in terms of language-specific factors, only; the possible influence of further influencing factors thus is not taken into consideration.

However, much of this criticism must then be directed towards contemporary research, too. Therefore, Mendenhall should be credited for having established an empirical basis for word length research, and for having initiated a line of

research which continues to be relevant still today. Particularly the last point mentioned above, leads to the next period in the history of word length studies. As can be seen, no attempt was made by Mendenhall to find a formal (mathematical) model, which might be able to describe (or rather, theoretically model) the frequency distribution. As a consequence, no objective comparison between empirical and theoretical distributions has been possible.

In this respect, the work of a number of researchers whose work has only recently and, in fact, only partially been appreciated adequately, is of utmost importance. These scholars have proposed particular frequency distribution models, on the one hand, and they have developed methods to test the goodness of the results obtained. Initially, most scholars have (implicitly or explicitly) shared the assumption that there might be one overall model which is able to represent a general theory of word length; more recently, ideas have been developed assuming that there might rather be some kind of general organizational principle, on the basis of which various specific models may be derived.

The present treatment concentrates on the rise and development of such models. It goes without saying that without empirical data, such a discussion would be as useless as the development of theoretical models. Consequently, the following presentation, in addition to discussing relevant theoretical models, will also try to present the results of empirical research. Studies of merely empirical orientation, without any attempt to arrive at some generalization, will not be mentioned, however – this deliberate concentration on theory may be an important explanation as to why some quite important studies of empirical orientation will be absent from the following discussion.

The first models were discussed as early as in the late 1940s. Research then concentrated on two models: the Poisson distribution, and the geometric distribution, on the other. Later, from the mid-1950s onwards, in particular the Poisson distribution was submitted to a number of modifications and generalizations, and this shall be discussed in detail below. The first model to be discussed at some length, here, is the geometric distribution which was suggested to be an adequate model by Elderton in 1949.

2. The Geometric Distribution (Elderton 1949)

In his article "A Few Statistics on the Length of English Words" (1949), English statistician Sir William P. Elderton (1877–1962), who had published a book on *Frequency-Curves and Correlation* some decades before (London 1906), studied the frequency of word lengths in passages from English writers, among them Gray, Macaulay, Shakespeare, and others.

As opposed to Mendenhall, Elderton measured word length in the number of syllables, not letters, per word. Furthermore, in addition to merely counting the frequencies of the individual word length classes, and representing them in

graphical form, Elderton undertook an attempt to find a statistical model for theoretically describing the distributions under investigation. His assumption was that the frequency distributions might follow the geometric distribution.

It seems reasonable to take a closer look at this suggestion, since, historically speaking, this was the first attempt ever made to arrive at a mathematical description of a word length frequency distribution. Where are zero-syllable words, i.e., if class $x = 0$ is not empty ($P_0 \neq 0$), the geometric distribution takes the following form (2.1):

$$P_x = p \cdot q^x \ , \quad x = 0, 1, 2, \dots \ , \quad 0 < q < 1 \ , p = 1 - q \ . \tag{2.1}$$

If class $x = 0$ is empty, however (i.e., if $P_0 = 0$), and the first class are one-syllable words (i.e., $P_1 \neq 0$) – then the geometric distribution looks as follows (2.2):

$$P_x = p \cdot q^{x-1} \ , \quad x = 1, 2, 3, \dots \tag{2.2}$$

Thus, generally speaking, for r-displaced distributions we may say:

$$P_x = p \cdot q^{x-r} \ , \quad x = r, r+1, r+2, \dots \tag{2.3}$$

Data given by Elderton (1949: 438) on the basis of letters by Gray, may serve as material to demonstrate the author's approach. Table 2.1 contains for each word length (x_i) the absolute frequencies (f_i), as given by Elderton, as well as the corresponding relative frequencies (p_i).[1]

There are various possibilities for estimating the parameter p of the geometric distribution when fitting the theoretical model to the empirical data. Elderton chose one of the standard options (at least of his times), which is based on the mean of the distribution:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{n} x_i \cdot f_i = \frac{7063}{5237} = 1.3487 \ .$$

Since, by way of the maximum likelihood method (or the log-likelihood method, respectively), it can be shown that, for $P_1 \neq 0$ ($x = 1, 2, 3, \dots$), p is the reciprocal of the mean, i.e. $p = 1/\bar{x}$; therefore, the calculation is as follows:

$$\hat{p} = 1/\bar{x} = 1/1.3487 = 0.7415$$

and

$$\hat{q} = 1 - p = 1 - 0.7415 = 0.2585.$$

[1] In his tables, Elderton added the data for these frequencies in per mille, and on this basis he then calculated the theoretical frequencies by fitting the geometric distribution to them. For reasons of exactness, only the raw data will be used in the following presentation and discussion of Elderton's data.

Table 2.1: Word Length Frequencies for English Letters by Th. Gray (Elderton 1949)

Number of syllables	Frequency of x-syllable words	
(x_i)	(f_i)	(p_i)
1	3987	0.7613
2	831	0.1587
3	281	0.0537
4	121	0.0231
5	15	0.0029
6	2	0.0004

In Elderton's English data, which are represented in Table 2.1, there are no zero-syllable words ($P_0 = 0$); we are thus concerned with a 1-displaced distribution. Therefore, formula (2.2) is to be applied. We thus obtain:

$$P_1 = P(X = 1) = 0.7415 \cdot 0.2585^{1-1} = 0.7415$$
$$P_2 = P(X = 2) = 0.7415 \cdot 0.2585^{2-1} = 0.1917 \text{ etc.}$$

Based on these probabilities, the theoretical frequencies can easily be calculated:

$$NP_1 = 5237 \cdot 0.7415 = 3883.08$$
$$NP_2 = 5237 \cdot 0.1917 = 1003.89 \text{ etc.}$$

The theoretical data, obtained by fitting the geometric distribution[2] to the empirical data from Table 2.1, are represented in Table 2.2 (cf. p. 22).

According to Elderton (1949: 442), the results obtained show that "the distributions [...] are not sufficiently near to geometrical progressions to be so described". Figure 2.3 (cf. p. 22) presents a comparison between the empirical data and the theoretical results, obtained by fitting the geometrical distribution to them (given in percentages). An inspection of this figure shows that Elderton's intuitive impression that the geometrical distribution is no adequate model to be fitted to the empirical data in a convincing manner, cannot clearly be corroborated.

[2] As compared to the calculations above, the theoretical frequencies slightly differ, due to rounding effects; additionally, for reasons not known, the results provided by Elderton (1949: 442) himself slightly differ from the results presented here, obtained by the method described by him.

Table 2.2: Fitting the Geometric Distribution to English Word Length Frequencies (Elderton 1949)

x_i	NP_i	P_i
1	3883.08	0.7415
2	1003.89	0.1917
3	259.54	0.0496
4	67.10	0.0128
5	17.35	0.0033
6	4.48	0.0009

As was rather usual in his time, Elderton did not run any statistical procedure to confirm his intuitive impression, i.e., to test the goodness of fit. Later, it would become a standard procedure to at least calculate a Pearson χ^2-goodness-of-fit value in order to test the adequacy of the theoretical model. Given this later development, it seems reasonable to re-analyze the result for Elderton's data in this respect.

Pearson's χ^2 is calculated by way of formula (2.4):

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - NP_i)^2}{E_i} \,. \tag{2.4}$$

In formula (2.4), k is number of classes, f_i is the observed frequency of a given class, and NP_i is the absolute theoretical frequency. For the data represented above, with $k = 6$ classes, we thus obtain $\chi^2 = 79.33$. The statistical significance of this χ^2 value depends on the degrees of freedom ($d.f.$), which

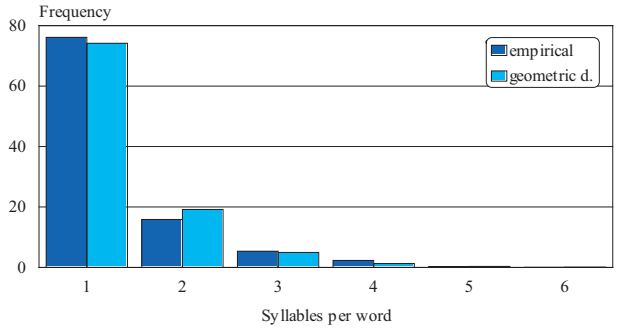


Figure 2.3: Empirical and Theoretical Word Length Frequencies (Elderton 1949)

in turn, are calculated with regard to the number of classes (k) minus 1, and the number of parameters (a) involved in the theoretical estimation: $d.f. = k-a-1$. Thus, with $d.f. = 6 - 2 = 4$ the χ^2 value obtained for Elderton's data can be interpreted in terms of a very poor fit indeed, since $p(\chi^2) < 0.001$.

However, it is a well-known fact that the value of χ^2 grows in a linear fashion with an increase of the sample size. Therefore, the larger a sample, the more likely the deviations tend to be statistically significant. Since linguistic samples tend to be rather larger, various suggestions have been brought forth as to a standardization of χ^2 scores. Thus, in contemporary linguistics, the discrepancy coefficient (C), which is easily calculated as $C = \chi^2/N$, has met general acceptance. The discrepancy coefficient, has the additional advantage that it is not dependent on degrees of freedom: in related studies, one speaks of a good fit for $C < 0.02$, and of a very good fit for $C < 0.01$.

In case of Elderton's data, we thus obtain a discrepancy coefficient of $C = 79.33/5237 = 0.015$; ultimately, this can be regarded to be an acceptable fit. Historically speaking, one should very much appreciate Elderton's early attempt to find an overall model for word length frequencies. What is problematic about his approach is not so much that his attempt was only partly successful for some English texts; rather, it is the fact that the geometrical distribution is adequate to describe monotonously decreasing distributions only. And although Elderton's data are exactly of this kind, word length frequencies from many other languages usually do not tend to display this specific shape.

Nevertheless, the geometric distribution has always attracted researchers' attention. Some decades later, Merkytė (1972), for example, discussed the geometric distribution with regard to its possible relevance for word length frequencies. Analyzing randomly chosen lexical material from a Lithuanian dictionary, he found differences as to the distribution of root words and words with affixes. As a first result, Merkytė (1972: 131) argued in favor of the notion "that the distribution of syllables in the roots is described by a geometric law", as a simple special case of the negative binomial distribution (for $k = 1$).

As an empirical test shows, the geometric distribution indeed turns out to be a good model. Since the data for the root words are given completely, the results given by Merkytė (1972: 128) are presented in Table 2.3 (p. 24).

As opposed to the root words, Merkytė found empirical evidence in agreement with the assumption that words with affixes follow a binomial distribution, i.e.

$$P_x = \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} \ , \quad x = 0, 1, \ldots n \ , \quad 0 < p < 1, q = 1 - p \ . \quad (2.5)$$

Unfortunately, no data are given for the words with affixes; rather, the author confines himself to theoretical ruminations on why the binomial distribution might be an adequate model. As a result, Merkytė (1972: 131) arrives at the

Table 2.3: Fitting the Geometric Distribution to Word Length Frequencies of Lithuanian Root Words (Merkytė 1972)

x_i	f_i	NP_i
1	525	518
2	116	135
3	48	34
4	9	9
5	2	2

hypothesis that the distribution of words is likely to be characterized as a "composition of geometrical and binomial laws".

In order to test his hypothesis, he gives, by way of an example, the relative frequencies of a list of dictionary words taken from a Lithuanian-French dictionary, represented in Table 2.4. Since the absolute sample size ($N = 25036$) is given as well, the absolute frequencies can easily be reconstructed as in Table 2.4.

Merkytė's combination of these two distributions results in the convolution of both for $x = 1, \ldots n$, and the geometric alone for $x = n+1, n+2, \ldots$; with a slight correction of Merkytė's presentation, it can be written as represented in formula (2.6):

$$P_x = \begin{cases} \sum_{i=0}^{x-1} \binom{n}{i} \alpha^i \beta^{n-i} p q^{x-i-1} & \text{for } x \leq n \\ \left(1 - \sum_{j=1}^{n} P_j\right) p q^{x-n-1} & \text{for } x > n \ . \end{cases} \tag{2.6}$$

Here, q is estimated as $\hat{q} = 1/\bar{x}_2$, where $\bar{x}_2$ is the mean word length of the sample's second part, i.e. its tail ($x > n$), and $\hat{p} = 1 - \hat{q}$. Parameter β, in turn, is estimated as $\hat{\beta} = (\bar{x} - \bar{x}_2)/n$, with $\hat{\alpha} = 1 - \hat{\beta}$.

The whole sample is thus arbitrarily divided into two portions, assuming that at a particular point of the data, there is a rupture in the material. With regard to the data presented in Table 2.4, Merkytė suggests $n = 3$ to be the crucial point. The approach as a whole thus implies that word length frequency would not be explained as an organic process, regulated by one overall mechanism, but as being organized by two different, overlapping mechanisms.

In fact, this is a major theoretical problem: Given one accepts the suggested separation of different word types – i.e., words with and without affixes – as a relevant explanation, the combination of both word types (i.e., the complete

Table 2.4: Theoretical Word Length Frequencies for Lithuanian Words: Merkytė-Geometric, Binomial and Conway-Maxwell-Poisson Distributions

		(Merkytė)	(Binomial)	(CMP)
x_i	f_i		NP_i	
1	3609	3734.09	3966.55	3346.98
2	9398	9147.28	8836.30	9544.32
3	7969	8144.84	7873.87	7965.80
4	3183	3232.87	3508.13	3240.21
5	752	651.59	781.51	791.50
6	125	125.31	69.64	147.19
	C	0.0012	0.0058	0.0012

material) does not, however, necessarily need to follow a composition of both individual distributions. Yet, the fitting of the Merkytė geometric distribution leads to convincing results: although the χ^2 value of $\chi^2 = 31.05$ is not really good ($p < 0.001$ for $d.f. = 3$), the corresponding discrepancy coefficient $C = 0.0012$ proves the fit to be excellent.[3] The results are represented in the first two columns of Table 2.4.

As a re-analysis of Merkytė's data shows, the geometric distribution cannot, of course, be a good model due to the lack of monotonous decrease in the data. However, the standard binomial distribution can be fitted to the data with quite some success: although the χ^2 value of $\chi^2 = 144.34$ is far from being satisfactory, resulting in $p < 0.001$ (with $d.f. = 3$), the corresponding discrepancy coefficient $C = 0.0058$ turns out be extremely good and proves the binomial distribution to be a possible model as well. The fact that the Merkytė geometric distribution turns out to be a better model as compared to the ordinary binomial distribution, is no wonder since after all, with its three parameters (α, p, n), the Merkytė geometric distribution has one parameter more than the latter.

Yet, this raises the question whether a unique, common model might not be able to model the Lithuanian data from Table 2.4. In fact, as the re-analysis shows, there is such a model which may very well be fitted to the data; we are concerned, here, with the Conway-Maxwell-Poisson (cf. Wimmer/Altmann 1999: 103), a standard model for word length frequencies, which, in its 1-displaced form, has the following shape:

[3] In fact, the re-analysis led to slightly different results; most likely, this is due to the fact that the data reconstruction on the basis of the relative frequencies implies minor deviations from the original raw data.

$$P_x = \frac{a^{x-1}}{(x-1)!^b T_1} , \quad x = 1, 2, 3, \dots , \quad T_1 = \sum_{j=1}^{\infty} \frac{a^j}{(j!)^b} . \tag{2.7}$$

Since this model will be discussed in detail below, and embedded in a broader theoretical framework (cf. p. 77), we will confine ourselves here to a demonstration of its good fitting results, represented in Table 2.4. As can be seen, the fitting results are almost identical as compared to Merkytės specific convolution of the geometric and binomial distributions, although the Conway-Maxwell-Poisson distribution has only two, not three parameters. What is more important, however, is the fact that, in the case of the Conway-Maxwell-Poisson distribution, no separate treatment of two more or less arbitrarily divided parts of the whole sample is necessary, so that in this case, the generation of word length follows one common mechanism.

With this in mind, it seems worthwhile to turn back to the historical background of the 1940s, and to discuss the work of Čebanov (1947), who, independent of and almost simultaneously with Elderton, discussed an alternative model of word length frequency distributions, suggesting the 1-displaced Poisson distribution to be of relevance.

3. The 1-Displaced Poisson Distribution (Čebanov 1947)

Sergej Grigor'evič Čebanov (1897–1966) was a Russian military doctor from Sankt Petersburg.[4] His linguistic interests, to our knowledge, mainly concentrated on the process of language development. He considered "the distribution of words according to the number of syllables" to be "one of the fundamental statistical characteristics of language structures", which, according to him, exhibits "considerable stability throughout a single text, or in several closely related texts, and even within a given language group" (Čebanov 1947: 99).

As Čebanov reports, he investigated as many as 127 different languages and vulgar dialects of the Indo-European family, over a period of 20 years. In his above-mentioned article – as far as we know, no other work of his on this topic has ever been published – Čebanov presented selected data from these studies, e.g., from High German, Iranian, Sanskrit, Old Irish, Old French, Russian, Greek, etc.

Searching a general model for the distribution of word length frequencies, Čebanov's starting expectation was a specific relation between the mean word length $\bar{x}$ of the text under consideration, and the relative frequencies p_i of the individual word length classes. In the next step, given the mean of the

[4] For a short biographical sketch of Čebanov see Best/Čebanov (2001).

distribution, Čebanov assumed the 1-displaced Poisson distribution to be an adequate model for his data. The Poisson distribution can be described as

$$P_x = \frac{e^{-a} \cdot a^x}{x!} \quad x = 0, 1, 2, \ldots \tag{2.8}$$

Since the support of (2.8) is $x = 0, 1, 2, \ldots$ with $a \geq 0$, and since there are no zero-syllable words in Čebanov's data, we are concerned with the 1-displaced Poisson distribution, which consequently takes the following shape:

$$P_x = \frac{e^{-a} \cdot a^{x-1}}{(x-1)!} \quad x = 1, 2, 3, \ldots \tag{2.9}$$

Čebanov (1947: 101) presented the data of twelve texts from different languages (or dialects). By way of an example, his approach will be demonstrated here, with reference to three texts. Two of these texts were studied in detail by Čebanov (1947: 102) himself: the High German text *Parzival*, and the Low Frankish text *Heliand*; the third text chosen here, by way of example, is a passage from Lev N. Tolstoj's *Vojna i mir* [War and Peace]. These data shall be additionally analyzed here because they are a good example for showing that word length frequencies do not necessarily imply a monotonously decreasing profile (cf. class $x = 2$) – it will be remembered that this was a major problem for the geometric distribution which failed be an adequate overall model (see above). The absolute frequencies (f_i), as presented by Čebanov (1947: 101), as well as the corresponding relative frequencies (p_i), are represented in Table 2.5 for all three texts.

Table 2.5: Relative Word Length Frequencies of Three Different Texts (Čebanov 1947)

Number of	*Parzival*		*Heliand*		*Vojna i mir*	
syllables (x_i)	f_i	p_i	f_i	p_i	f_i	p_i
1	1823	0.6280	1572	0.4693	466	0.2826
2	849	0.2925	1229	0.3669	541	0.3281
3	194	0.0668	452	0.1349	391	0.2371
4	37	0.0127	83	0.0248	172	0.1043
5			14	0.0042	64	0.0388
6					15	0.0091
$\sum$	2903		3350		1698	

As can be seen from Figure 2.4, all three distributions clearly seem to differ from each other in their shape; particularly the *Vojna i mir* passage, displaying a peak at two-syllable words, differs from the two others.

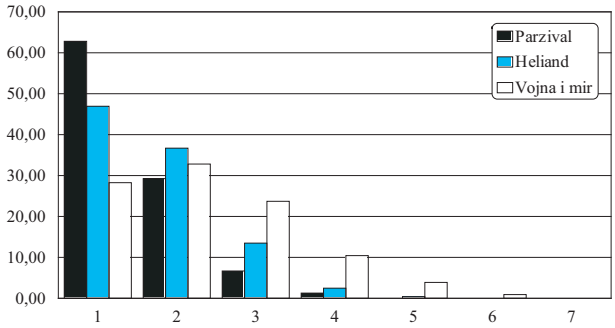


Figure 2.4: Empirical Word Length Frequencies of Three Texts (Čebanov 1947)

How then, did the Poisson distribution in its 1-displaced form fit? Let us demonstrate this with reference to the data from *Parzival* in Table 2.5. Since the mean in this text is $\bar{x} = 1.4643$, with $\hat{a} = \bar{x} - 1$ and referring to formula (2.9) for the 1-displaced Poisson distribution, we thus obtain

$$P_x = \frac{e^{-(1.4643-1)} \cdot (1.4643-1)^{x-1}}{(x-1)!} \,. \tag{2.10}$$

Thus, for $x = 1$ and $x = 2$, we obtain

$$P_1 = \frac{e^{-0.4643} \cdot 0.4643^0}{0!} = \frac{2.7183^{-0.4643} \cdot 1}{1} = 0.6285$$

$$P_2 = \frac{e^{-0.4643} \cdot 0.4643^1}{1!} = 2.7183^{-0.4643} \cdot 0.4643 = 0.2918$$

Correspondingly, for $x = 1$ and $x = 2$, we receive the following theoretical frequencies:

$$NP_1 = 2903 \cdot 0.6285 = 1824.54$$

$$NP_2 = 2903 \cdot 0.2918 = 847.10$$

Table 2.6 contains the results of fitting the 1-displaced Poisson distribution to the empirical data of the three texts, or text passages, also represented in Table 2.5 above.[5]

Whereas Elderton, in his analyses, did not run any statistical procedures to statistically test the adequacy of the proposed model, Čebanov did so. Well

[5] As compared to the calculations above, the theoretical frequencies slightly differ, due to rounding effects. For reasons not known, the results also differ as compared to the data provided by Čebanov (1947: 102), obtained by the method described above.

Table 2.6: Fitting the 1-Displaced Poisson Distribution to Word Length Frequencies (Čebanov 1947)

Number of	*Parzival*		*Heliand*		*Vojna i mir*	
syllables (x_i)	f_i	NP_i	f_i	NP_i	f_i	NP_i
1	1823	1824.67	1572	1618.01	466	442.29
2	849	847.28	1229	1177.53	541	582.04
3	194	196.72	452	428.48	391	382.97
4	37	30.45	83	103.94	172	167.99
5			14	18.91	64	55.27
6					15	14.55
$\sum$	2903		3350		1698	

aware of A.A. Markov's (1924) *caveat*, that "complete coincidence of figures cannot be expected in investigations of this kind, where theory is associated with experiment", Čebanov (1947: 101) calculated χ^2 goodness-of-fit values. As a result, Čebanov (ibd.) arrived at the conclusion that the χ^2 values "show good agreement in some cases and considerable departure in others." Let us follow his argumentation step by step, based on the three texts mentioned above.

For *Parzival*, with $k = 4$ classes, we obtain $\chi^2 = 1.45$. This χ^2 value can be interpreted in terms of a very good fit, since $p(\chi^2) = 0.48$ $(d.f. = 2)$.[6] Whereas the 1-displaced Poisson distribution thus turns out to be a good model for *Parzival*, Čebanov interprets the results for *Heliand* not to be: here, the value is $\chi^2 = 10.35$, which, indeed, is a significantly worse, though still acceptable result ($p = 0.016$ for $d.f. = 3$).[7]

Interestingly enough, the 1-displaced Poisson distribution would also turn out to be a good model for the passage from Tolstoj's *Vojna i mir* (not analyzed in detail by Čebanov himself), with a value of $\chi^2 = 5.82$ ($p = 0.213$ for $d.f. = 4$).

On the whole, Čebanov (1947: 101) arrives at the conclusion that the theoretical results "show good agreement in some cases and considerable departure in others." This partly pessimistic estimation has to be corrected however. In fact, Čebanov's (1947: 102) interpretation clearly contradicts the intuitive impression one gets from an optical inspection of Figure 2.5: as can be seen, $P_i(a)$, represented for $i = 1, 2, 3$, indeed seems to be "determined all but completely"

[6] Čebanov (1947: 102) himself reports a value of $\chi^2 = 0.43$ which he interprets to be a good result.

[7] Čebanov (1947: 102) reports a value of $\chi^2 = 13.32$ and, not indicating any degrees of freedom, interprets this result to be a clear deviation from expectation.

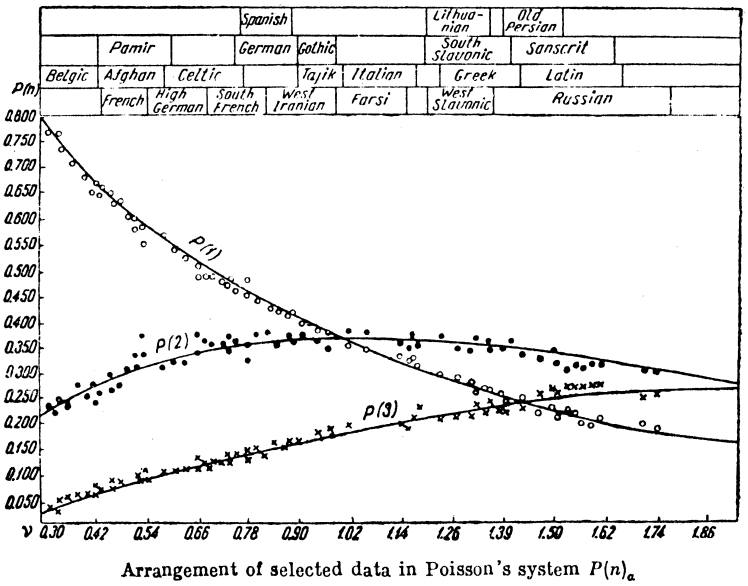


Figure 2.5: The 1-Displaced Poisson Distribution as a Word Length Frequency Distribution Model (Čebanov 1947)

by the mean of the text under consideration (ibd., 101). In Figure 2.5, Poisson's $P_i(a)$ can be seen on the horizontal, the relative frequencies for p_i on the vertical axis).

The good fit of the 1-displaced Poisson distribution may also be proven by way of a re-analysis of Čebanov's data, calculating the discrepancy values C (see above). Given that in case of all three texts mentioned and analyzed above, we are concerned with relatively large samples ($N = 2903$ for *Parzival*, $N = 1649$ for *Heliand*, and $N = 1698$ for the *Vojna i mir* passage). In fact, the result is $C < 0.01$ in all three cases.[8] In other words: what we have here are excellent fits, in all three cases, which can be clearly seen in the graphical illustration of Figure 2.6 (p. 31).

Unfortunately, Čebanov's work was consigned to oblivion for a long time. If at all, reference to his work was mainly made by some Soviet scholars, who, terming the 1-displaced Poisson distribution "Čebanov-Fucks distribution", would later place him on a par with German physician Wilhelm Fucks. As is well known, Fucks and his followers would also, and independently of

[8] As a corresponding re-analysis of the twelve data sets given by Čebanov (1947: 101) shows, C values are $C < 0.02$ in all cases, and they are even $C < 0.01$ in two thirds of the cases.

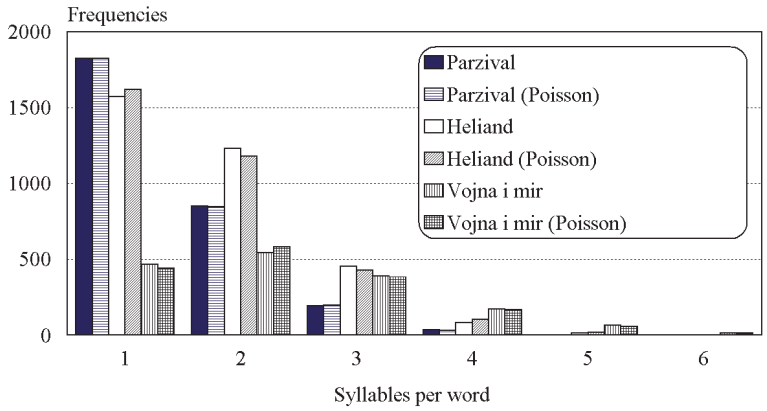


Figure 2.6: Fitting the 1-Displaced Poisson Distribution to Three Text Segments (Čebanov 1947)

Čebanov's work, favor the 1-displaced Poisson distribution to be an important model, in the late 1950s. Before presenting Fucks' work in detail, it is necessary to discuss another approach, which also has its roots in the 1940s.

4. The Lognormal Distribution

A different approach to theoretically model word length distributions was pursued mainly in the late 1950s and early 1960s by scholars such as Gustav Herdan (1958, 1966), René Moreau (1963), and others.

As opposed to the approaches thus far discussed, these authors did not try to find a discrete distribution model; rather, they worked with continuous models, mainly the so-called lognormal model.

Herdan was not the first to promote this idea with regard to language. Before him, Williams (1939, 1956) had applied it to the study of sentence length frequencies, arguing in favor of the notion that the frequency with which sentences of a particular length occur, are lognormally distributed. This assumption was brought forth, based on the observation that sentence length or word length frequencies do not seem to follow a normal distribution; hence, the idea of lognormality was promoted. Later, the idea of word length frequencies being lognormally distributed was only rarely picked up, such as for example by Russian scholar Piotrovskij and colleagues (Piotrovskij et al. 1977: 202ff.; cf. 1985: 278ff.).

Generally speaking, the theoretical background of this assumption can be characterized as follows: the frequency distribution of linguistic units (as of other units occurring in nature and culture) often tends to display a right-sided asymmetry, i.e., the corresponding frequency distribution displays a positive

skewness. One of the theoretical reasons for this can be seen in the fact that the variable in question cannot go beyond (or remain below) a particular limit; since it is thus characterized by a one-sided limitation in variation, the distribution cannot be adequately approximated by the normal distribution.

Particularly when a distribution is limited by the value 0 to the left side, one suspects to obtain fairly normally distributed variables by logarithmic transformations: as a result, the interval between 0 and 1 is transformed into $-\infty$ to 0. In other words: the left part of the distribution is stretched, and at the same time, the right part is compressed. The crucial idea of lognormality thus implies that a given random variable X follows a lognormal distribution if the random variable $Y = \log(X)$ is normally distributed. Given the probability density function for the normal distribution as in (2.11),

$$y = f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty \tag{2.11}$$

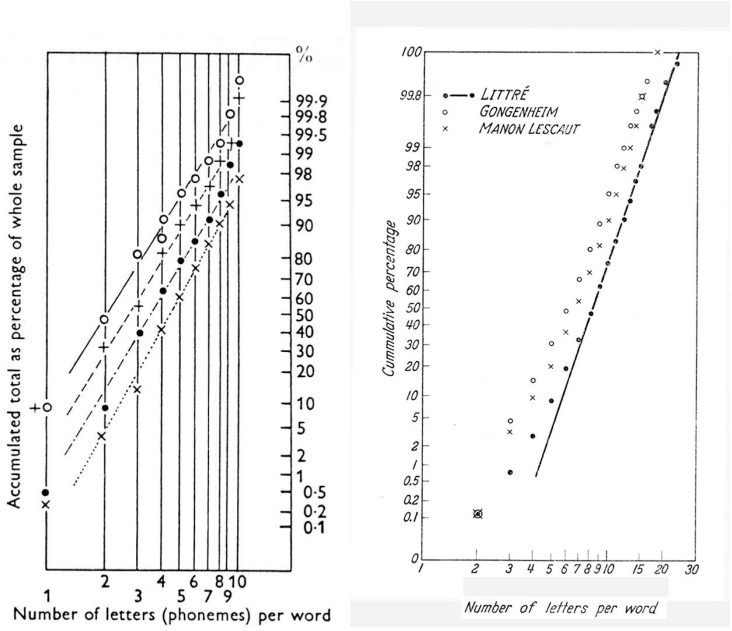
one thus obtains the probability density function for the lognormal distribution in equation (2.12):

$$y = f(x) = \frac{1}{\sigma \cdot x \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}, \quad 0 < x < \infty \tag{2.12}$$

Herdan based his first analyses of word length studies on data by Dewey (1923) and French et al. (1930). These two studies contain data on word length frequencies, the former 78,633 words of written English, the latter 76,054 words of spoken English. Thus, Herdan had the opportunity to do comparative analyses of word length frequencies measured in letters and phonemes. In order to test his hypothesis as to the lognormality of the frequency distribution, Herdan (1966: 224) confined himself to graphical techniques only. The most widely applied method in his time was the use of probability grids, with a logarithmically divided abscissa (x-axis) and the cumulative frequencies on the ordinate (y-axis). If the resulting graph showed a more or less straight line, one regarded a lognormal distribution to be proven.

As can be seen from Figure 2.7, the result seems to be quite convincing, both for letters and phonemes. In his later monograph on *The Advanced Theory of Language as Choice and Chance*, Herdan (1966: 201ff.) similarly analyzed French data samples, taken from analyses by Moreau (1963). The latter had analyzed several French samples, among them the three picked up by Herdan in Figure 2.7:

1. 3,204 vocabulary entries from George Gougenheim's *Dictionnaire fondamental de la langue française*,
2. 76,918 entries from Émile Littré's *Dictionnaire de la langue française*
3. 6,151 vocubulary items from the *Histoire de Chevalier des Grieux et de Manon Lescaut* by the Abbé Prévost.



(a) Herdan (1958: 224) (b) Herdan (1966: 203)

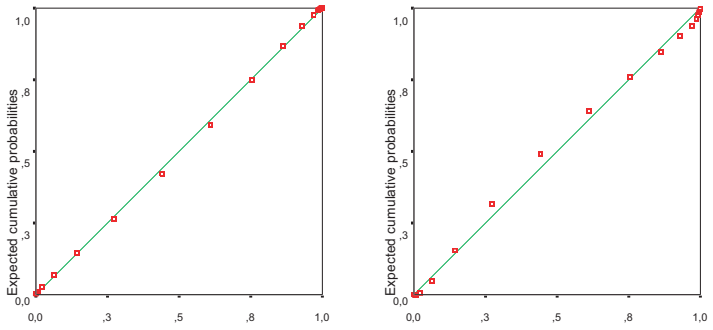
Figure 2.7: Word Length Frequencies on a Lognormal Probability Grid (Herdan 1958/66)

The corresponding graph is reproduced in Figure 2.7. Again, for Herdan (1966: 203), the inspection of the graph "shows quite a satisfactory linearity [...], which admits the conclusion of lognormality of the distribution."

In this context, Herdan discusses Moreau's (1961, 1963) introduction of a third parameter (V_0) into the lognormal model, ultimately causing a displacement of the distribution; as can be seen, $\theta \cdot \log k$ is a mere re-parametrization of σ – cf. (2.12).

$$f(x) = \frac{1}{(\theta \log k) \cdot (x + V_0) \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{\log(x+V_0)-\log k}{\theta \log k}\right)^2} . \tag{2.13}$$

Herdan considered this extension not to be necessary. In his book, he offered theoretical arguments for the lognormal distribution to be an adequate model (Herdan 1966: 204). These arguments are in line with the general characteristics of the lognormal distribution, in which the random variables are considered to influence each other in a multiplying manner, whereas the normal distribution is characterized by the additive interplay of the variables (the variables thus being considered to be independent of one another).



(a) Normal distribution (b) Lognormal distribution)

Figure 2.8: *P-P* Plots for Fitting the Normal and Lognormal Distributions to Word Length Frequencies in Abbé Prévost's *Manon Lescaut*

However, Herdan did not do any comparative analyses as to the efficiency of the normal or the lognormal distribution, neither graphically nor statistically. Therefore, both procedures shall be presented here, by way of a re-analysis of the original data.

As far as graphical procedures are concerned, probability grids have been replaced by so-called *P-P* plots, today, which also show the cumulative proportions of a given variable and should result in a linear rise in case of normal distribution. By way of an example, Figure 2.8 represents the *P-P* plots for *Manon Lescaut*, tested for normal and lognormal distribution. It can clearly be seen that there are quite some deviations for the lognormal distribution (cf. Figure2.8(b)). What is even more important, however, is the fact that the deviations are clearly less expressed for the normal distribution (cf. Figure2.8(a)). Although this can, in fact, be shown for all three data samples mentioned above, we will concentrate on a statistical analysis of these observations.

Table 2.7 contains the relevant Kolmogorov-Smirnov values (KS) and the corresponding p-values with the given degrees of freedom ($d.f.$) for all three samples, both for the normal and the logarithmized values. Additionally, values for skewness (γ_1) and kurtosis (γ_2) are given, so that the effect of the logarithmic manipulation of the data can easily be seen.

As can clearly be seen, the deviations both from the normal and the lognormal distributions are highly significant in all cases. Furthermore, differences between normal and lognormal are minimal; in case of *Manon Lescaut*, the lognormal distribution is even worse than the normal distribution.

Table 2.7: Statistical Comparison of Normal and Lognormal Distributions for Three French Texts (Herdan 1966)

		KS	df	p	γ_1	γ_2
Manon Lescaut	normal distr.	0.105	6151	< 0.0001	0.30	0.22
	lognormal d.	0.135			−0.89	1.83
Littré	normal distr.	0.108	76917	< 0.0001	0.53	0.49
	lognormal d.	0.103			−0.47	0.68
Gougenheim	normal distr.	0.121	3204	< 0.0001	0.80	2.06
	lognormal d.	0.126			−0.55	1.12

The same holds true, by the way, for the above-mentioned data presented by Piotrovskij et al. (1985: 283). The authors analyzed a German technical text of 1,000 words and found, as they claimed, "a clear concordance between the empirical distribution and the lognormal distribution of the random variable". As a justification of their claim they referred to a graphical representation of empirical and theoretical values, only; however, they additionally maintained that the assumed concordance may easily and strongly be proven by way of Kolmogorov's criterium (ibd., 281).

As a re-analysis of the data shows, this claim may not be upheld, however (cf. table 2.8):

Table 2.8: Statistical Comparison of Normal and Lognormal Distribution for German Data (Piotrovskij 1985)

	KS	df	p	γ_1	γ_2
normal distr.	0.12	1000	0.0001	0.81	0.20
lognormal d.	0.08			−0.25	−0.52

As in the case of Herdan's analyses, the effect of the logarithmic transformation can easily be deduced from the values for γ_1 and γ_2 (i.e., for skewness and kurtosis). Also, the deviation from the normal distribution is highly significant ($p < 0.001$). However, as can be seen the deviation from the lognormal distribution is highly significant as well, and, strictly speaking, even greater compared to the normal distribution.

In summary, one can thus say that neither the normal distribution nor the lognormal distribution model turns out to be adequate in praxis. With regard to

this negative finding, one may add the result of a further re-analysis, saying that in case of all three data samples discussed by Herdan, the binomial distribution can very well be fitted to the empirical data, with $0.006 \leq C \leq 0.009$. No such fit is possible in the case of Piotrovskij's data, however, which may be due to the fact that the space was considered to be part of a word.

Incidently, Michel (1982) arrived at the very same conclusion, in an extensive study on Old and New Bulgarian, as well as Old and New Greek material. He tested the adequacy of the lognormal distribution for the word length frequencies of the above-mentioned material on two different premises, basing his calculation of word length both on the number of letters per word, and on the number of syllables per word. As a result, Michel (1982: 198) arrived at the conclusion "that the fitting fails completely".[9]

One can thus say that there is overwhelming negative empirical evidence which proves that the lognormal distribution is no adequate model for word length frequencies of various languages. Additionally, and this is even more important in the given context, one must state that there are also major theoretical problems which arise in the context of the (log-)normal distribution as a possible model for word length frequencies:

a. the approximation of continuous models to discrete data;

b. the doubtful dependence of the variables, due to the multiplying effect of variables within the lognormal model;

c. the manipulation of the initial data by way of logarithmic transformations.

With this in mind, let us return to discrete models. The next historical step in the history of word length studies were the important theoretical and empirical analyses by Wilhelm Fucks, a German physician, whose theoretical models turned out to be of utmost importance in the 1950s and 1960s.

5. The Fucks Generalized Poisson Distribution

5.1 The Background

As mentioned previously, the 1-displaced Poisson distribution had been suggested by S.G. Čebanov in the late 1940s. Interestingly enough, some years later the very same model – i.e., the 1-displaced Poisson distribution – was also favored by German physicist Wilhelm Fucks (1955a,b, 1956b). Completely independent of Čebanov, without knowing the latter's work, and based on completely different theoretical assumptions, Fucks arrived at similar conclusions to

[9] Michel also tested the adequacy of the 1-displaced Poisson distribution (see below, p. 46).

those of Čebanov some years before. However, Fucks' work was much more influential than was Čebanov's, and it was Fucks rather than Čebanov, who would later be credited for having established the 1-displaced Poisson distribution as a standard model for word length frequency distributions.

When Fucks described the 1-displaced Poisson distribution and applied it to his linguistic data, he considered it to be "a mathematical law, thus far not known in mathematical statistics" (Fucks 1957: 34). In fact, he initially derived it from a merely mathematical perspective (Fucks 1955c); in his application of it to the study of language(s) and language style(s), he then considered it to be the "general law of word-formation" (1955a: 88, 1957: 34), or, more exactly, as the "mathematical law of the process of word-formation from syllables for all those languages, which form their words from syllables" (Fucks 1955b: 209).

In fact, Fucks' suggestion was the most important model discussed from the 1950s until the late 1970s; having the 1-displaced Poisson distribution in mind, one used to refer to it as "the Fucks model". Only in Russia, one should later speak of the "Čebanov-Fucks distribution" (e.g., Piotrovskij et al. 1977: 190ff.; cf. Piotrowski et al. 1985: 256ff.), thus adequately honoring the pioneering work of Čebanov, too.

There was one major difference between Čebanov's and Fucks' approaches, however: this difference has to be seen in the fact that Fucks' approach was based on a more general theoretical model, the 1-displaced Poisson distribution being only one of its special cases (see below). Furthermore, Fucks, in a number of studies, developed many important ideas on the general functioning not only of language, but of other human sign systems, too. This general perspective as to the "mathematical analysis of language, music, or other results of human cultural activity" (Fucks 1960: 452), which is best expressed in Fucks' (1968) monograph *Nach allen Regeln der Kunst*, cannot be dealt with in detail, here, where our focus is on the history of word length studies.

5.2 The General Approach

Ultimately, Fucks' general model can be considered to be an extension of the Poisson distribution; specifically, we are concerned with a particularly weighted Poisson distribution. These weights are termed $\varepsilon_k - \varepsilon_{k+1}$, k indicating the number of components to be analyzed.

In its most general form, this weighting generalization results in the following formula (2.14):

$$p_i = P(X = i) = e^{-\lambda} \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \cdot \frac{\lambda^{i-k}}{(i-k)!} \,. \tag{2.14}$$

Here, the random variable X denotes the number of syllables per word, i.e. $X = i, i = 0, 1, 2, 3, \ldots, I$. The probability that a given word has i syllables,

is $p_i = P(X = i)$, with $\sum_{i=0}^{I} p_i = 0$, $\lambda = \mu - \varepsilon'$, $\varepsilon' = \sum_{k=1}^{\infty} \varepsilon_k$ and $\mu = E(X)$. The parameters of the distribution $\{\varepsilon_k\}$ are called the ε-spectrum. For (2.14), there are a number of conditions postulated by Fucks which must be fulfilled:

(a) From the necessity that $\varepsilon_k - \varepsilon_{k+1} \geq 0$ it follows that $\varepsilon_{k+1} \leq \varepsilon_k$;

(b) Since the sum of all weights equals 1, we have
$1 = \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) = \sum_{k=0}^{\infty} \varepsilon_k - \sum_{k=0}^{\infty} \varepsilon_{k+1} = \varepsilon_0$; it follows that $\varepsilon_0 = 1$.

Finally, from (a) and (b) it follows

(c) $1 = \varepsilon_0 \geq \varepsilon_1 \geq \varepsilon_2 \geq \varepsilon_3 \geq \ldots \geq \varepsilon_k \geq \varepsilon_{k+1} \ldots$

As can be seen from equation (2.14), the so-called "generalized Fucks distribution" includes both the standard Poisson distribution (2.8) and the 1-displaced Poisson distribution (2.9) as two of its special cases. Assuming that $\varepsilon_0 = 1$, and $\varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_k = 0$ – one obtains the standard Poisson distribution (2.8):

$$p_i = e^{-\lambda} \cdot \frac{\lambda^i}{i!} \qquad i = 0, 1, 2, \ldots$$

Likewise, for $\varepsilon_0 = \varepsilon_1 = 1$, and $\varepsilon_2 = \varepsilon_3 = \ldots = \varepsilon_k = 0$, one obtains the 1-displaced Poisson distribution (2.9) (cf. p. 27):

$$p_i = e^{-\lambda} \cdot \frac{\lambda^{i-1}}{(i-1)!}, \quad i = 1, 2, \ldots$$

As was already mentioned above, the only model which met general acceptance was the 1-displaced Poisson distribution. More often than not, Fucks himself applied the 1-displaced Poisson distribution without referring to his general model, and this may be one more reason why it has often (though rather incorrectly) been assumed to be "the Fucks distribution". In other words: In spite of the overwhelming number of analyses presented by Fucks in the 1950s and 1960s, and irrespective of the broad acceptance of the 1-displaced Poisson distribution as an important model for word length studies, Fucks' generalization as described above can only be found in very few of his works (e.g., Fucks 1956a,b).

It is no wonder, then, that the generalized model has practically not been discussed. Interestingly enough, however, several scholars of East European background became familiar with Fucks' concept, and they not only discussed it at some length, but also applied it to specific data. It seems most reasonable to assume that this rather strange circumstance is due to the Russian translation of Fucks' 1956b paper (cf. Fucks 1957).

Before turning to the East European reception of Fucks' model, resulting not only in its application, but also in some modification of it, let us first discuss some of the results obtained by Fucks in his own application of the 1-displaced Poisson distribution to linguistic data.

5.3 The 1-Displaced Poisson Distribution as a Special Case of Fucks' Generalization of the Poisson Distribution

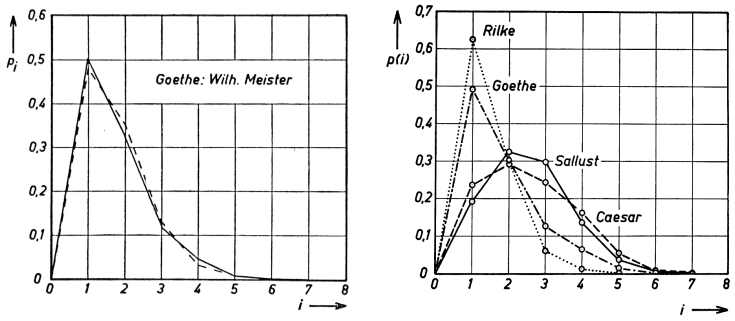
In his inspiring works, Fucks applied the 1-displaced Poisson distribution on different levels of linguistic and textual analysis: on the one hand, he analyzed single texts, but he also studied word length frequency distribution in text corpora, both from one and the same language and across languages. Thus, his application of the 1-displaced Poisson distribution included studies on (1) the individual style of single authors, as well as on (2) texts from different authors either (2.1) of one and the same language or (2.2) of different languages.

As an example of the study of individual texts, Figure 2.9(a) from Fucks (1956b: 208) may serve. It shows the results of Fucks' analysis of Goethe's *Wilhelm Meister*: on the horizontal x-axis, the number of syllables per word (termed i by Fucks) are indicated, on the vertical y-axis the relative frequency of each word length class (p_i) can be seen. As can be seen from the dotted line in Figure 2.9(a), the fitting of the 1-displaced Poisson distribution seems to result in extremely good theoretical values.

As to a comparison of two German authors, Rilke and Goethe, on the one hand, and two Latin authors, Sallust and Caesar, on the other, Figure 2.9(b) may serve. It gives rise to the impression that word length frequency may be characteristic of a specific author's style, rather than of specific texts. Again, the fitting of the 1-displaced Poisson distribution seems to be convincing.

There can be no doubt about the value of Fucks' studies, and still today, they contain many inspiring ideas which deserve to be further pursued. Yet, in re-analyzing his works, there remains at least one major problem: Fucks gives many characteristics of the specific distributions, starting from mean values and standard deviations up to the central moments, entropy etc. Yet, there are hardly ever any raw data given in his texts, a fact which makes it impossible to check the results at which he arrived. Thus, one is forced to believe in the goodness of his fittings on the basis of his graphical impressions, only; and this drawback is further enhanced by the fact that there are no procedures which are applied to test the goodness of his fitting the 1-displaced Poisson distribution. Ultimately, therefore, Fucks' works cannot but serve as a starting point for new studies which would have to replicate his results.

There is only one instance where Fucks presents at least the *relative*, though not the absolute frequencies of particular distributions in detail. This is when he presents the results of a comparison of texts from nine different languages – eight



(a) Goethe's *Wilhelm Meister* (b) German and Latin authors

Figure 2.9: Fitting the 1-Displaced Poisson Distribution to German and Latin Text Segments (Fucks 1956)

natural languages, and one artificial (cf. Fucks 1955a: 85ff.). The results for each language are based on what Fucks (1955a: 84) considered to be "representative cross-sections of written documents" of the given languages.

The relative frequencies are reproduced in Table 2.9 which, in addition to the relative frequency of each word length class (measured in syllables per word), also contains the mean ($\bar{x}$), as well as the entropy (H) for each language, the latter being calculated by way of formula (2.15):

$$H = -\sum_{i=1}^{n} p_i \ln p_i \ . \tag{2.15}$$

Unfortunately, quite a number of errors can be found in Fuck' original table, both as to the calculated values of $\bar{x}$ and H; therefore, the data in Table 2.9 represent the corrected results which one obtains on the basis of the relative frequencies given by Fucks and formula (2.15). We will come back to these data throughout the following discussion, using them as exemplifying material. Being well aware of the fact that for each of the languages we are concerned with mixed data, we can ignore this fact, and see the data as a representation of a maximally broad spectrum of different empirical distributions which may be subjected to empirical testing.

Figure 2.10 (p. **??**) illustrates the frequency distributions, based on the relative frequencies of the word length classes for each language. The figure is taken from Fucks (1955a: 85), since the errors in the calculation concern only $\bar{x}$ and H and are not relevant here. According to Fucks' interpretation, all shapes fall into one and the same profile, except for Arabic; as a reason for this, Fucks

Table 2.9: Relative Frequencies, Mean Word Length, and Entropy for Different Languages (Fucks 1955)

	English	**German**	**Esperanto**	**Arabic**	**Greek**
1	0.7152	0.5560	0.4040	0.2270	0.3760
2	0.1940	0.3080	0.3610	0.4970	0.3210
3	0.0680	0.0938	0.1770	0.2239	0.1680
4	0.0160	0.0335	0.0476	0.0506	0.0889
5	0.0056	0.0071	0.0082	0.0017	0.0346
6	0.0012	0.0014	0.0011	–	0.0083
7	–	0.0002	–	–	0.0007
8	–	0.0001	–	–	–
$\bar{x}$	1.4064	1.6333	1.8971	2.1106	2.1053
H	0.3665	0.4655	0.5352	0.5129	0.6118
	Japanese	**Russian**	**Latin**	**Turkish**	
1	0.3620	0.3390	0.2420	0.1880	
2	0.3440	0.3030	0.3210	0.3784	
3	0.1780	0.2140	0.2870	0.2704	
4	0.0868	0.0975	0.1168	0.1208	
5	0.0232	0.0358	0.0282	0.0360	
6	0.0124	0.0101	0.0055	0.0056	
7	0.0040	0.0015	0.0007	0.0004	
8	0.0004	0.0003	0.0002	0.0004	
9	0.0004	–	–	–	
$\bar{x}$	2.1325	2.2268	2.3894	2.4588	
H	0.6172	0.6355	0.6311	0.6279	

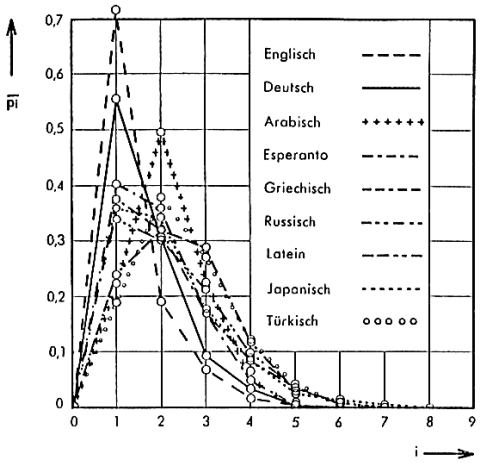


Figure 2.10: Relative Frequencies of Word Lengths in Eight Natural and One Artificial Languages (Fucks 1955)

assumed that the number of texts analyzed in this language might not have been sufficient.

As was mentioned above, Fucks did not, as was not unusual at his time, calculate any tests as to the significance of the goodness of his fits. It seems that Fucks (1955a: 101) was very well aware of the problems using the χ^2-goodness-of-fit test for this purpose, since he explicitly emphasized that, "for particular mathematical reasons", his data are "not particularly adequate" for the application of the χ^2 test.

The problem behind Fucks' assumption might be the fact that the χ^2 value increases in a linear way with an increase of sample size; therefore, results are more likely to display significant differences for larger samples, which is almost always the case in linguistic studies. As was mentioned above (cf. p. 23), the problem is nowadays avoided by calculating the discrepancy coefficient $C = \chi^2/N$, which is not dependent on the degrees of freedom. We may thus easily, by way of a re-analysis, calculate C for the data given by Fucks, in order to statistically test the goodness-of-fit of the 1-displaced Poisson distribution; in order to do so, we simply have to create "artificial" samples of ca. 10,000 each, by multiplying the relative frequencies with 10,000.

Remembering that fitting is considered to be good in case of $0.01 < C < 0.02$ and very good in case of $C < 0.01$, one has to admit that fitting the 1-displaced Poisson distribution to Fucks' data from different languages is not really convincing (see Table 2.10): strictly speaking, it turns out to be adequate only for an artificial language, Esperanto, and must be discarded as an overall valid model.

Table 2.10: Discrepancy Coefficient C as a Result of Fitting the 1-Displaced Poisson Distribution to Different Languages (Fucks 1955)

	English	**German**	**Esperanto**	**Arabic**	**Greek**
C (1-par.)	0.0903	0.0186	0.0023	0.1071	0.0328
	Japanese	**Russian**	**Latin**	**Turkish**	
C (1-par.)	0.0380	0.0208	0.0181	0.0231	

It is difficult to say whether the observed failure is due to the fact that the data for each of the languages originated from text mixtures (and not from individual texts), or if there are other reasons. Still, Fucks and many followers of his pursued the idea of the 1-displaced Poisson distribution as the most adequate model for word length frequencies.

Although Fucks did not calculate any statistics to test the goodness of fit (which in fact many people would not do still today), one must do him justice and point out that he tried to go another way to empirically prove the adequacy of his findings: knowing the values of $\bar{x}$ and H for each language, Fucks graphically illustrated their relationship and interdependency. Figure 2.11 shows the results, with $\bar{x}$ on the horizontal x-axis, and H on the vertical y-axis; the data are based on the corrected values from Table 2.9.

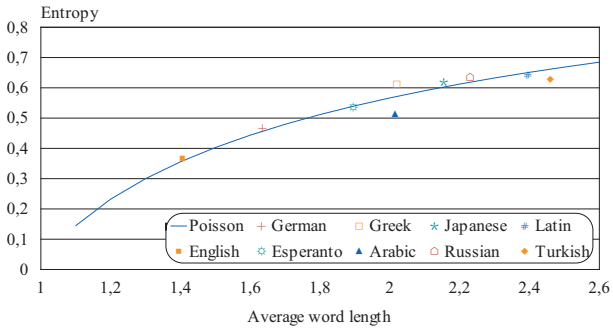


Figure 2.11: Entropy as a Function of Mean Word Length (Fucks 1955a)

Additionally, Fucks calculated the entropy of the theoretical distribution, estimating $\hat{a}$ as $\bar{x}$; these values can easily be obtained by formula (2.8) (cf. p. 27), and they are reproduced below in Table 2.11. Thus, one arrives at the curve in Figure 2.11, representing the Poisson distribution (cf. Fucks 1955a: 85). As can

Table 2.11: Empirical and Theoretical Entropy for Nine Word Length Frequency Distributions (Fucks)

$\bar{x}$	$H\ [y]$	$\hat{H}\ [\hat{y}]$
1.4064	0.3665	0.3590
1.6333	0.4655	0.4563
1.8971	0.5352	0.5392
2.1032	0.5129	0.5913
2.1106	0.6118	0.5917
2.1325	0.6172	0.6030
2.2268	0.6355	0.6184
2.3894	0.6311	0.6498
2.4588	0.6279	0.6614

be seen with Fucks (1955a: 88, 1960: 458f.), the theoretical distribution "represents the values found in natural texts very well". In other words: evaluating his results, Fucks once again confined himself to merely visual impressions, as he did in the case of the frequency probability distribution. And again, it would have been easy to run such a statistical test, calculating the coefficient of determination (R^2) in order to test the adequacy of the theoretical curve obtained.

Let us shortly discuss this procedure: in a nonlinear regression model, R^2 represents that part of the variance of the variable y, which can be explained by variable x. There are quite a number of more or less divergent formulae to calculate R^2 (cf. Grotjahn 1982), which result in partly significant differences. Usually, the following formula (2.16) is taken:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \ . \tag{2.16}$$

With $0 \leq R^2 \leq 1$, one can say that the greater R^2, the better the theoretical fit. In order to calculate R^2, we thus consider $\bar{x}$ to be the independent variable x, and H to be the dependent variable y. Thus, for each empirical x_i, we need both the empirical values (y_i) and the theoretical values ($\widehat{y}_i$) which can be obtained by formula (2.8), and which are represented in Table 2.11. Based on these results, we can now easily calculate R^2, with $\bar{y} = \overline{H[y]}$ (cf. Table 2.11), as

$$R^2 = 1 - \frac{0.0097}{0.0704} = 0.8768 \ . \tag{2.17}$$

As can be seen, the fit can be regarded to be relatively good.[10] This result is not particularly influenced by the fit for Arabic, which, according to Fucks, deviates from the other languages. In fact, the value for R^2 hardly changes if one, following Fucks' argumentation, eliminates the data for Arabic: under this condition, the determination coefficient would result in $R^2 = 0.8763$.

Still, there remains a major theoretical problem with the specific method chosen by Fucks in trying to prove the adequacy of the 1-displaced Poisson distribution: this problem is related to the method itself, i.e., in establishing a relation between $\bar{x}$ and H. Taking a second look at formula (2.15), one can easily see that the entropy of a frequency distribution is ultimately based on p_i, only; p_i, however, in case of the Poisson distribution, is based on parameter a in formula (2.8), which is nothing but the mean $\bar{x}$ of the distribution! In other words, due to the fact that the Poisson distribution is mainly shaped by the mean of the distribution, Fucks arrives at a tautological statement, relating the mean $\bar{x}$ of the Poisson distribution to its entropy H.

To summarize, one has thus to draw an important conclusion: Due to the fact that Fucks did not apply any suitable statistics to test the goodness of fit for the 1-displaced Poisson distribution, he could not come to the point of explicitly stating that this model may be adequate in some cases, but is not acceptable as a general standard model. Still, Fucks' suggestions had an enormous influence on the study of word length frequencies, particularly in the 1960's. Most of these subsequent studies concentrated on the 1-displaced Poisson distribution, as suggested by Fucks.

In fact, work on the Poisson distribution is by no means a matter of the past. Rather, subsequent to Fucks' (and of course Čebanov's) pioneering work on the Poisson distribution, there have been frequent studies discussing and trying to fit the 1-displaced Poisson distribution to linguistic data, with and without reference to the previous achievements.

No reference to Fucks (or Čebanov) is made, for example, by Rothschild (1986) in his study on English dictionary material. Rothschild's initial discussion of previous approaches to word length frequencies, both continuous and discrete, was particularly stimulated by his disapproval of Bagnold's (1983) assumption that word length distributions are not Gaussian, but skew (hyperbolic or double exponential) distributions. Discussing and testing various distribution models, Rothschild did not find any one of the models he tested to be adequate. This holds true for the (1-displaced) Poisson distribution, as well, which, according to Rothschild (1986: 317), "fails on a formal χ^2-test". Nevertheless, he considered it to be "the most promising candidate" (ibd., 321) – quite obviously, *faute de mieux*...

[10] Calculating the determination coefficient with the data given by Fucks himself results in $R^2 = 0.8569$.

As opposed to Rothschild, Michel (1982), in his above-mentioned study of Old and New Bulgarian and Greek material (cf. p. 36), explicitly referred to Fucks' work on the Poisson distribution. As was said above, Michel first found the lognormal distribution to be a completely inadequate model. He then tested the 1-displaced Poisson distribution and obtained negative results as well: although fitting the Poisson distribution led to better results as compared to the lognormal distribution, word length in his data turned out not to be Poisson distributed, either (Michel 1982: 199f.)

Finally, Grotjahn (1982) whose work will be discussed in more detail below (cf. p. 61ff.), explicitly discussed Fucks' work on the 1-displaced Poisson distribution, being able to show under which empirical conditions it is likely to be an adequate model, and when it is prone to fail. He too, however, did not discuss the 1-displaced Poisson distribution as a special case of Fucks' generalized Poisson model.

It seems reasonable, therefore, to follow Fucks' own line of thinking. In doing so, let us first direct our attention to the 2-parameter model suggested by him, and then to his 3-parameter model.

5.4 The (1-Displaced) Dacey-Poisson Distribution as a 2-Parameter Special Case of Fucks' Generalization of the Poisson Distribution

It has been pointed out in the preceding section that for $\varepsilon_0 = 1$, and $\varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_k = 0$ the standard Poisson distribution (2.8) is obtained from formula (2.14). Likewise, for $\varepsilon_0 = \varepsilon_1 = 1$, and $\varepsilon_2 = \varepsilon_3 = \ldots = \varepsilon_k = 0$, one obtains the 1-displaced Poisson distribution (2.9), which has been discussed above (cf. p. 27). In either case, the result is a 1-parameter model in which only λ has to be estimated.

In a similar way, two related 2-parameter distributions can be derived from the general model (2.14) under the following circumstances: In case of $\varepsilon_0 = 1,\ \varepsilon_1 \neq 0$, and $\varepsilon_k = 0$ for $k \geq 2$, one obtains the so-called Dacey-Poisson distribution (cf. Wimmer/Altmann 1999: 111), replacing ε_1 by α:

$$p_i = (1-\alpha) \cdot \frac{e^{-\lambda}\lambda^i}{i!} + \alpha \cdot \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!}\,, \quad i = 0, 1, 2, \ldots \tag{2.18}$$

with $\lambda = \mu - \alpha$. Here, in addition to λ, a second parameter ($\varepsilon_1 = \alpha$) has to be estimated, e.g., as $\hat{\alpha} = \sqrt{\bar{x} - m_2}$.

Similarly, for $\varepsilon_0 = \varepsilon_1 = 1,\ \varepsilon_2 \neq 0,$ and $\varepsilon_k = 0$ for $k \geq 3$, one obtains a model which has become known as the 1-displaced Dacey-Poisson distribution (2.19), replacing ε_2 by α:

$$p_i = (1-\alpha) \cdot \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!} + \alpha \cdot \frac{e^{-\lambda}\lambda^{i-2}}{(i-2)!}\,, \quad i = 1, 2, \ldots \tag{2.19}$$

with $\lambda = (\mu - \alpha) - 1$; in this case, α can be estimated as $\hat{\alpha} = \sqrt{\bar{x} - 1 - m_2}$. It is exactly the latter distribution (2.19) which has been discussed by Fucks as another special case of his generalized Poisson model, though not under this name. Fucks has not systematically studied its relevance; still, it might be tempting to see what kind of results are yielded by this distribution for the data already analyzed above (cf. Table 2.10). Table 2.12 (which additionally contains the dispersion quotient d to be explained below) represents the values of the discrepancy coefficient C as a result of a corresponding re-analysis.

Table 2.12: Discrepancy Coefficient C as a Result of Fitting the 1-Displaced Dacey-Poisson Distribution to Different Languages (Fucks 1955)

	English	**German**	**Esperanto**	**Arabic**	**Greek**
C (2-par.)	—	—	0.0019	0.0077	—
d	1.3890	1.1751	0.9511	0.5964	1.2179
	Japanese	**Russian**	**Latin**	**Turkish**	
C (2-par.)	—	—	0.0149	0.0021	
d	1.2319	1.1591	0.8704	0.8015	

As can be seen from Table 2.12, in some cases, the results are slightly better as compared to the results obtained from fitting the 1-displaced Poisson distribution (cf. Table 2.10). However, in some cases no results can be obtained. The reason for this failure is the fact that the estimation of α as $\hat{\alpha} = \sqrt{\bar{x} - 1 - m_2}$ (see above) results in a negative root, obviously due to the fact that the estimate $\hat{\alpha}$ is not defined if $\bar{x} - 1 \leq m_2$.

Recently, Stadlober (2003) gave an explanation for this finding. Referring to Grotjahn's (1982) work, which will be discussed below (cf. p. 61ff.), Stadlober analyzed the theoretical scope of Fucks' 2-parameter model. Grotjahn's interest had been to find out under what conditions the 1-displaced Poisson distribution can be an adequate model for word length frequencies. Therefore, Grotjahn (1982) had suggested to calculate the quotient of dispersion δ, based on the theoretical values for a sample's mean (μ) and its variance (σ^2):

$$\delta = \frac{\sigma^2}{\mu} \; . \tag{2.20}$$

For r-displaced distributions, the corresponding equation is

$$\delta = \frac{\sigma^2}{\mu - r} \; , \tag{2.21}$$

r being the displacement parameter.

The coefficient δ can, of course, be calculated not only for theoretical frequencies, but also for empirical frequencies, then having the notation

$$d = \frac{m_2}{\bar{x} - r} \ . \tag{2.22}$$

Given both the empirical value of d and the value of δ, one can easily test the goodness of fitting the Poisson distribution to empirical data, by calculating the deviation of d (based on the empirical data) from δ (as the theoretical value to be expected). Now, since, for the 1-displaced Poisson distribution, the variance $Var(X) = \sigma^2 = \mu - 1$, we have

$$\delta = \frac{\mu - 1}{\mu - 1} = 1 \ .$$

The logical consequence arising from the fact that for the Poisson distribution, $\delta = 1$, is that the latter can be an adequate model only as long as $d \approx 1$ in an empirical sample. Now, based on these considerations, Stadlober (2003) explored the theoretical dispersion quotient δ for the Fucks 2-parameter distribution (2.19), discussed above. Since here, $Var(X) = \mu - 1 - \varepsilon_2^2$, it turns out that $\delta \leq 1$; this means that this 2-parameter model is likely to be inadequate as a theoretical model for empirical samples with $d > 1$.

As in the case of the 1-displaced Poisson distribution, one has thus to acknowledge that the Fucks 2-parameter (1-displaced Dacey-Poisson) distribution is an adequate theoretical model only for a specific type of empirical distributions. This leads to the question whether the Fucks 3-parameter distribution is more adequate as an overall model.

5.5 The 3-Parameter Fucks-Distribution as a Special Case of Fucks' Generalization of the Poisson Distribution

In the above sections, the 1-displaced Poisson distribution and the 1-displaced Dacey-Poisson distribution were derived as two special cases of the Fucks Generalized Poisson distribution as described in (2.14). In the first case, the ε-spectrum had the form $\varepsilon_0 = \varepsilon_1 = 1, \varepsilon_k = 0$ for $k \geq 2$, and in the second case $\varepsilon_0 = \varepsilon_1 = 1, \ \varepsilon_2 = \alpha, \varepsilon_k = 0$ for $k \geq 3$.

Now, in the case of the 3-parameter model, ε_2 and ε_3 have to be estimated, the whole ε-spectrum having the form: $\varepsilon_0 = \varepsilon_1 = 1$, $\varepsilon_2 = \alpha, \varepsilon_3 = \beta, \varepsilon_k = 0$ for $k \geq 4$, resulting in the following model:

$$p_i = e^{-(\mu - 1 - \alpha - \beta)} \cdot \sum_{k=1}^{3} (\varepsilon_k - \varepsilon_{i+1}) \frac{(\mu - 1 - \alpha - \beta)^{i-k}}{(i-k)!} \tag{2.23}$$

Replacing $\lambda = mu - \alpha - \beta$, the probability mass function has the following form:

$$p_1 = e^{-\lambda} \cdot (1 - \alpha)$$

$$p_2 = e^{-\lambda} \cdot [(1 - \alpha) \cdot \lambda + (\alpha - \beta)]$$

$$p_i = e^{-\lambda} \left[(1 - \alpha) \frac{\lambda^{i-1}}{(i-1)!} + (\alpha - \beta) \frac{\lambda^{i-2}}{(i-2)!} + \beta \frac{\lambda^{i-3}}{(i-3)!} \right], \quad i \geq 3$$

As to the estimation of $\varepsilon_2 = \alpha$ and $\varepsilon_3 = \beta$, Fucks (1956a: 13) suggested calculating them by reference to the second and third central moments (μ_2 and μ_3). It would lead too far, here, to go into details, as far as their derivation is concerned. Still, the resulting 2×2-system of equations shall be quoted:

(a) $\mu_2 = \mu_1 - 1 - (\alpha + \beta)^2 + 2\beta$

(b) $\mu_3 = \mu_3 = \mu_1 + 2(1+\alpha+\beta)^3 - 3(1+\alpha+\beta)^2 - 6(\alpha+\beta)(\alpha+2\beta) + 6\beta$

As can be seen, the solution of this equation system – which can be mathematically simplified (cf. Antić/Grzybek/Stadlober 2005a) – involves a cubic equation. Consequently, three solutions are obtained, not all of which must necessarily be real solutions. For each real solution the values for $\varepsilon_2 = \alpha$ and $\varepsilon_3 = \beta$ have to be estimated (which is easily done by computer programs today, as opposed to in Fucks' time).[11]

Before further going into details of this estimation, let us remember that there are two important conditions as to the two parameters:

(a) $\varepsilon_2 = \alpha \leq 1$ and $\varepsilon_3 = \beta \leq 1$,

(b) $\varepsilon_2 = \alpha \geq \beta = \varepsilon_3$.

With this in mind, let us once again analyze the data of Table 2.9, this time fitting Fucks' 3-parameter model. The results obtained can be seen in Table 2.13; results not meeting the two conditions mentioned above, are marked as $\varnothing$.

It can clearly be seen that in some cases, quite reasonably, the results for the 3-parameter model are better, as compared to those of the two models discussed above. One can also see that the 3-parameter model may be an appropriate model for empirical distributions in which $d > 1$ (which was the decisive problem for the two models described above): thus, in the Russian sample, for example, where $d = 1.1591$, the discrepancy coefficient is $C = 0.0005$. However, as the results for German and Japanese data (with $d = 1.1751$ and $d = 1.2319$, respectively) show, d does not seem to play the crucial role in case of the 3-parameter model.

[11] In addition to a detailed mathematical reconstruction of Fucks' ≪Theory of Word Formation≫, Antić/Grzybek/Stadlober (2005b) have tested the efficiency of Fucks' model in empirical research.

Table 2.13: Discrepancy Coefficient C as a Result of Fitting the Fucks 3-Parameter Poisson Distribution to Different Languages (Fucks 1955)

	English	**German**	**Esperanto**	**Arabic**	**Greek**
C	∅	∅	0.00004	0.0021	∅
$\hat{\varepsilon}_2$	—	—	0.3933	0.5463	—
$\hat{\varepsilon}_3$	—	—	0.0995	-0.1402	—
d	1.3890	1.1751	0.9511	0.5964	1.2179
	Japanese	**Russian**	**Latin**	**Turkish**	
C	∅	0.0005	0.0003	0.0023	
$\hat{\varepsilon}_2$	—	0.2083	0.5728	0.6164	
$\hat{\varepsilon}_3$	—	0.1686	0.2416	0.1452	
d	1.2319	1.1591	0.8704	0.8015	

In fact, as Antić/Grzybek/Stadlober (2005a) show, the conditions for the Fucks 3-parameter model to be appropriate are slightly different. The details need not be discussed here; it may suffice to say that it is ultimately the difference $M = \bar{x} - m_2$, i.e. the difference between the mean of the empirical distribution ($\bar{x}$) and its variance (m_2). One thus obtains the following two conditions:

1. The sum $a = \hat{\varepsilon}_2 + \hat{\varepsilon}_3 = \hat{\alpha} + \hat{\beta}$ must be in a particular interval:

$$a_i \in \left[\frac{1 - \sqrt{4M - 3}}{2}, \frac{1 + \sqrt{4M - 3}}{2} \right] \ , \ i = 1, 2, 3$$

   Thus, there are two interval limits a_1 and a_2:

$$a_{i1} = \frac{1 - \sqrt{4M - 3}}{2} \qquad \text{and} \qquad a_{i2} = \frac{1 + \sqrt{4M - 3}}{2} \ .$$

2. In order to be $a \in \mathbb{R}$, the root $4M - 3$ must be positive, i.e. $4M - 3 \geq 0$; therefore, $M = \bar{x} - m_2 \geq 0.75$.

From the results represented in Table 2.14 (p. 51) it can clearly be seen why, in four of the nine cases, the results are not as desired: there are a number of violations, which are responsible for the failure of Fucks' 3-parameter model. These violations can be of two kinds:

a. As soon as $M < 0.75$, the definition of the interval limits of a_1 and a_2 involves a negative root – this is the case with the Japanese data, for example;

b. Even if the first condition is met with $M \geq 0.75$, fitting the Fucks 3-parameter model may fail, if the condition $a_{i1} < a < a_{i2}$ is not fulfilled – this can be seen in the case of the English, German, and Greek data.

Table 2.14: Violations of the Conditions for Fucks' 3-Parameter Model

	English	**German**	**Esperanto**	**Arabic**	**Greek**
C	∅	∅	< 0.01	< 0.01	∅
$\hat{\varepsilon}_2$	—	—	0.3933	0.5463	—
$\hat{\varepsilon}_3$	—	—	0.0995	-0.1402	—
$a = \hat{\varepsilon}_2 + \hat{\varepsilon}_3$	-0.0882	-0.1037	0.4929	0.4061	0.2799
a_{i1}	0.1968	0.1270	-0.0421	-0.3338	0.4108
a_{i2}	0.8032	0.8730	1.0421	1.3338	0.5892
$a_{i1} < a < a_{i2}$	–	–	✓	✓	–
$\bar{x}$	1.4064	1.6333	1.8971	2.1032	2.1106
m_2	0.5645	0.7442	0.8532	0.6579	1.3526
$M = \bar{x} - m_2$	0.8420	0.8891	1.0438	1.4453	0.7580
$M \geq 0.75$	✓	✓	✓	✓	✓
	Japanese	**Russian**	**Latin**	**Turkish**	
C	∅	< 0.01	< 0.01	< 0.01	
$\hat{\varepsilon}_2$	—	0.2083	0.5728	0.6164	
$\hat{\varepsilon}_3$	—	0.1686	0.2416	0.1452	
$a = \hat{\varepsilon}_2 + \hat{\varepsilon}_3$	-0.1798	0.3769	0.8144	0.7616	
a_{i1}	ℂ	0.2659	-0.1558	-0.2346	
a_{i1}	ℂ	0.7341	1.1558	1.2346	
$a_{i1} < a < a_{i2}$	–	✓	✓	✓	
$\bar{x}$	2.1325	2.2268	2.3894	2.4588	
m_2	1.3952	1.4220	1.2093	1.1692	
$M = \bar{x} - m_2$	0.7374	0.8048	1.1800	1.2896	
$M \geq 0.75$	–	✓	✓	✓	

Fucks' 3-parameter model thus is adequate only for particular types of empirical distributions, and it can not serve as an overall model for language, not even for syllabic languages, as Fucks himself claimed. However, some of the problems met might be related to the specific way of estimating the parameters suggested by him, and this might be the reason why other authors following him tried to find alternative ways.

5.6 The Georgian Line: Cercvadze, Čikoidze, Cilosani, Gačečiladze

Quite early, three Georgian scholars, G.N. Cercvadze, G.B. Čikoidze, and T.G. Gačečiladze (1959), applied Fucks' ideas to Georgian linguistic material, mainly to phoneme frequencies and word length frequencies. Their study, which was translated into German as early as 1962, and which was later extended by Gačečiladze/Cilosani (1971), was originally inspired by the Russian translation of Fucks' English-language article ≪Mathematical Theory of Word Formation≫. Fucks' article, originally a contribution to the 1956 London Conference on *Information Theory*, had been published in England in 1956, and it was translated into Russian only one year later, in 1957. As opposed to most of his German papers, Fucks had discussed his generalization at some length in this English synopsis of his work, and this is likely to be the reason why his approach received much more attention among Russian-speaking scholars.

In fact, Cercvadze, Čikoidze, and Gačečiladze (1959) based their analyses on Fucks' generalization; the only thing different from Fucks' approach is their estimation of the two parameters ε_2 and ε_3 of Fucks 3-parameter model: as opposed to Fucks, they estimated ε_2 and ε_3 not with recourse to the central moments, but to the initial moments of the empirical distribution. The empirical central moment of the order r

$$m_r = \frac{1}{(N-1)} \sum_x (x - \bar{x})^r f_x$$

is an estimate of the r-th theoretical moment defined as

$$\mu_r = \sum_x (x - \mu)^r P_x.$$

As estimate for the theoretical initial moment of the order r

$$\mu'_r = \sum_x x^r P_x$$

serves the empirical r-th initial moment given as:

$$m'_r = \frac{1}{N}\sum_x x^r f_x \ .$$

Since it can be shown that central moments and initial moments can be transformed into each other, the results can be expected to be identical; still, the procedure of estimating is different.

We need not go into details, here, as far as the derivation of the Fucks distribution and its generating function is concerned (cf. Antić/Grzybek/Stadlober 2005a). Rather, it may suffice to name its first three initial moments, which are necessary for the equation system to be established, which, in turn, is needed for the estimation of ε_2 and ε_3. Thus, with

$$\sum_{k=1}^{\infty}\varepsilon_k = \varepsilon' \tag{2.24}$$

we have the first three initial moments of Fucks' distribution:

$$\begin{aligned}
\mu'_1 &= \mu \\
\mu'_2 &= \mu^2 + \mu - \varepsilon'^2 - 2\varepsilon' + 2\sum_{k=1}^{\infty} k\varepsilon_k \\
\mu'_3 &= \mu^3 + 3\mu^2 + \mu + 2\varepsilon'^3 + 3\varepsilon'^2 - \varepsilon' - 3\mu\varepsilon'^2 - 6\mu\varepsilon' + \\
&\quad + \sum_{k=0}^{\infty} k^3\left(\varepsilon_k - \varepsilon_{k+1}\right) + 6\left(\mu - \varepsilon'\right)\sum_{k=1}^{\infty} k\varepsilon_k
\end{aligned} \tag{2.25}$$

Now, replacing ε_2 with α, and ε_3 with β, we obtain the following system of equations:

(a) $\mu'_2 = \mu^2 + \mu - (1+\alpha+\beta)^2 - 2(1+\alpha+\beta) + 2(1+2\alpha+3\beta)$

(b) $\mu'_3 = \mu^3 + 3\mu^2 + \mu + 2(1+\alpha+\beta)^3 + 3(1-\mu)(1+\alpha+\beta)^2 + 6\alpha + 18\beta - 6\mu(1+\alpha+\beta) + 6(\mu - 1 - \alpha - \beta)(1+2\alpha+3\beta)$.

After the solution for α and β, we thus have the following probabilities:

$$\begin{aligned}
p_1 &= e^{-\lambda}\cdot(1-\alpha) \\
p_2 &= e^{-\lambda}\cdot[(1-\alpha)\cdot\lambda + (\alpha-\beta)] \\
p_i &= e^{-\lambda}\left[(1-\alpha)\frac{\lambda^{i-1}}{(i-1)!} + (\alpha-\beta)\frac{\lambda^{i-2}}{(i-2)!} + \beta\frac{\lambda^{i-3}}{(i-3)!}\right], \quad i \geq 3
\end{aligned}$$

with $\lambda = \mu - 1 - \alpha - \beta$.

As was said above, the results are identical as compared to those obtained by recourse to the central moments. Unfortunately, there are several mistakes in the authors' own formula; therefore, there is no sense in reproducing their results on their Georgian sample, here.[12]

Almost twenty years later, Russian scholars Piotrovskij, Bektaev, and Piotrovskja (1977: 193; cf. 1985: 261), would again refer to Fucks' generalized model. These authors quite rightly termed the above-mentioned 1-displaced Poisson distribution (2.9) the "Čebanov-Fucks distribution" (cf. p. 27). In addition to this, they mentioned the so-called "generalized Gačečiladze-Fucks distribution", which deserves some more attention here.

As was seen above, the 1959 paper by Cercvadze, Čikoidze, and Gačečiladze was based on Fuck's generalization of the Poisson distribution. Obviously, these authors indeed once again generalized the Fucks model, which is not inherent in the 1959 paper mentioned, but represented in an extension of it by Gačečiladze/Cilosani (1971). This extension contains an additional factor φ_{nu}, which is dependent on three parameters:

(a) the mean of the sample ($\bar{i}$),

(b) the relevant class i,

(c) the sum of all ε_ν, $A = \sum_{\nu=1}^{\infty} \varepsilon_\nu$ (termed ε' by Fucks).

As a result, the individual weights of the generalized Fucks distribution, defined as $(\varepsilon_k - \varepsilon_{k+1})$, are multiplied by the function φ_ν. Unfortunately, Gačečiladze/Cilosoni (1971: 114) do not explain the process by which φ_{nu} may be theoretically derived; they only present the final formula (2.26):

$$P_i = e^{-\left(\bar{i}-A\right)} \sum_{\nu=0}^{\infty} (\varepsilon_\nu - \varepsilon_{\nu+1}) \frac{(\lambda - A)^{i-\nu}}{(i-\nu)!} \varphi_\nu \left(A, \bar{i}, i\right) \tag{2.26}$$

Here, $\bar{i}$ is the mean of the sample, and $(\varepsilon_k - \varepsilon_{k+1})$ are the weighting factors. Unfortunately, Piotrovskij et al. (1977: 195), who term formula (2.26) the "Fucks-Gačečiladze distribution", also give no derivation for ϕ_ν. Assuming that φ_ν takes account of the contextual environment, they only refer to Fucks' 1955 *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen*. However, neither Fucks' generalization nor φ are mentioned in this work. Thus, as to the theoretical derivation of φ_ν, there are only sparse references by Gačečiladze/Cilosani (1971: 114) who mentioned some of their Georgian publications, which are scarcely available.

[12] In fact, in spite of, or rather due to their obvious calculation errors, the authors arrived at a solution for ε_2 and ε_3, which yields a good theoretical result; these values cannot be derived from the correct formula, however, and therefore must be considered to be a casual and accidental ad hoc solution.

Still, it can easily be seen that for $\phi_\nu \to 1$, one obtains the generalized Fucks distribution, which has also been discussed by some Polish authors.

5.7 Estimating the Fucks Distribution with First-Class Frequency (Bartkowiakowa/Gleichgewicht 1964/65)

Two Polish authors, Anna Bartkowiakowa and Bolesław Gleichgewicht (1964, 1965), also suggested an alternative way to estimate the two parameters ε_2 and ε_3 of Fucks' 3-parameter distribution. Based on the standard Poisson distribution, as represented in (2.27),

$$g_k = \frac{\lambda^k}{k!} e^{-\lambda} \ , \quad k = 0, 1, 2, \ldots \tag{2.27}$$

and referring to Fucks' (2.14) generalization of it, the authors reformulated the latter as seen in (2.28):

$$\begin{aligned} p_i &= \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) e^{-\lambda} \frac{\lambda^{i-k}}{(i-k)!} \\ &= \sum_{k=0}^{\infty} (\varepsilon_k - \varepsilon_{k+1}) \cdot g_{i-k} \ . \end{aligned} \tag{2.28}$$

Determining $\varepsilon_0 = \varepsilon_1 = 1$, and $\varepsilon_k = 0$ for $k > 3$, the two parameters $\varepsilon_2 \neq 0$ and $\varepsilon_3 \neq 0$ remain to be estimated on the basis of the empirical distribution. Based on these assumptions, the following special cases are obtained for (2.28):

$$\begin{aligned} p_1 &= (1 - \varepsilon_2) \cdot g_0 \\ p_2 &= (1 - \varepsilon_2) \cdot g_1 + (\varepsilon_2 - \varepsilon_3) \cdot g_0 \\ p_i &= (1 - \varepsilon_2) \cdot g_{i-1} + (\varepsilon_2 - \varepsilon_3) \cdot g_{i-2} + \varepsilon_3 \cdot g_{i-3} \ \text{ for } \ i \geq 3 \end{aligned}$$

with $\lambda = \mu - (1 + \varepsilon_2 + \varepsilon_3)$.

As to the estimation of ε_2 and ε_3, the authors did not set up an equation system on the basis of the second and third central moments (μ_2 and μ_3), as did Fucks, thus arriving at a cubic equation; rather, they first defined the portion of one-syllable words (p_1), and then modelled the whole distribution on that proportion. Thus, by way of a logarithmic transformation of $p_1 = (1 - \varepsilon_2) \cdot g_0$ in formula (2.28), one obtains the following sequence of transformations:

$$\begin{aligned} \ln \frac{p_1}{(1 - \varepsilon_2)} &= \ln g_0 \\ \ln \frac{p_1}{(1 - \varepsilon_2)} &= -\lambda \\ \ln \frac{p_1}{(1 - \varepsilon_2)} &= -\left[\mu - (1 + \varepsilon_2 + \varepsilon_3)\right] \ . \end{aligned}$$

Referring to the empirical distribution, a first equation for an equation system to be solved (see below) is thus gained from the first probability (p_1) of the empirical distribution:

$$\ln \frac{\hat{p}_1}{(1-\hat{\varepsilon}_2)} = -\left[\bar{x} - (1+\hat{\varepsilon}_2+\hat{\varepsilon}_3)\right] \tag{2.29}$$

The second equation for that system is then gained from the variance of the empirical distribution. Thus, one gets

$$\mu_2 = \mu - (1+\varepsilon_2+\varepsilon_3)^2 + 2\cdot(\varepsilon_2 + 2\cdot\varepsilon_3)$$

resulting in the second equation for the equation system to be established:

$$m_2 = \bar{x} - (1+\hat{\varepsilon}_2+\hat{\varepsilon}_3)^2 + 2\cdot(\hat{\varepsilon}_2 + 2\hat{\varepsilon}_3) \tag{2.30}$$

With the two equations (2.29) and (2.30), we thus have the following system of equations, adequate to arrive at a solution for ε_2 and ε_3:

$$\boxed{\begin{array}{l} \text{(a) } \ln \frac{\hat{p}_1}{(1-\hat{\varepsilon}_2)} = -\left[\bar{x} - (1+\hat{\varepsilon}_2+\hat{\varepsilon}_3)\right] \\ \\ \text{(b) } m_2 - \bar{x} = -(1+\hat{\varepsilon}_2+\hat{\varepsilon}_3)^2 + 2\,(\hat{\varepsilon}_2 + 2\hat{\varepsilon}_3) \end{array}}$$

Bartkowiakowa/Gleichgewicht (1964) not only theoretically presented this procedure to estimate ε_2 and ε_3; they also offered the results of empirical studies, which were meant to be a test of their model. These analyses comprised nine Polish literary texts, or segments of them, and the results of these analyses indeed proved their approach to be successful.

Table 2.15 contains the results: as can be seen, the discrepancy coefficient is $C < 0.01$ in all cases; furthermore, in six of the nine samples, the result is indeed better as compared to Fucks' original estimation.

For the sake of comparison, Table 2.15 also contains the results for the (1-displaced) Poisson and the (1-displaced) Dacey-Poisson distributions, which were calculated in a re-analysis of the raw data provided by the Polish authors. A closer look at these data shows that the Polish text samples are relatively homogeneous: for all texts, the dispersion quotient is in the interval $0.88 \leq d \leq 1.04$, and $0.95 \leq M \leq 1.09$.

Table 2.15: Fitting the Fucks 3-Parameter Model to Polish Data, with Parameter Estimation Based on First-Class Frequency

	1	**2**	**3**	**4**	**5**
$\bar{x}$	1.81	1.82	1.96	1.93	2.07
m_2	0.76	0.73	0.87	0.94	1.07
d	0.93	0.88	0.91	1.00	0.99
M	1.05	1.09	1.09	0.99	1.00
C (Poisson)	0.00420	0.00540	0.00370	0.00170	0.00520
C (Dacey-Poisson)	0.00250	0.00060	0.00200	∅	0.00531
$C\ (m_2, m_3)$	0.00240	0.00017	0.00226	0.00125	0.00085
$C\ (\hat{p}_1, m_2)$	0.00197	0.00043	0.00260	0.00194	0.00032

	6	**7**	**8**	**9**
$\bar{x}$	2.12	2.05	2.18	2.16
m_2	1.10	0.98	1.21	1.21
d	0.98	0.94	1.03	1.04
M	1.02	1.07	0.97	0.95
C (Poisson)	0.00810	0.00220	0.01360	0.00940
C (Dacey-Poisson)	0.00862	0.00145	∅	∅
$C\ (m_2, m_3)$	0.00084	0.00120	0.00344	0.00383
$C\ (\hat{p}_1, m_2)$	0.00030	0.00077	0.00216	0.00271

This raises the question in how far the procedure suggested by Bartkowiakowa/Gleichgewicht (1964) is able to improve the results for the nine different languages analyzed by Fucks himself (cf. Table 2.9, p. 41). Table 2.16 represents the corresponding results.

In summary, one may thus say that the procedure to estimate the two parameters ε_2 and ε_3, as suggested by Bartkowiakowa/Gleichgewicht (1964), may indeed, for particular samples, result in better fittings. However, they cannot overcome the overall limitations of Fucks' 3-parameter model, which have been discussed above.

Table 2.16: Fucks' 3-Parameter Model, with Parameter Estimation

	Esperanto	**Arabic**	**Russian**	**Latin**	**Turkish**
m_2, m_3					
$\hat{\varepsilon}_2$	0.3933	0.5463	0.2083	0.5728	0.6164
$\hat{\varepsilon}_3$	0.0995	-0.1402	0.1686	0.2416	0.1452
C	0.00004	0.0021	0.0005	0.0003	0.0023
$\hat{p}_1, m_2$					
$\hat{\varepsilon}_2$	0.3893	0.7148	0.2098	0.5744	0.6034
$\hat{\varepsilon}_3$	0.0957	0.1599	0.1695	0.2490	0.1090
C	0.00001	0.0042	0.0005	0.0003	0.0018

6. The Doubly Modified Poisson Distribution (Vranić/Matković 1965)

A different approach to modify the standard Poisson distribution has been suggested by Vranić/Matković (1965a,b). The authors analyzed Croatian data from two corpora, each consisting of several literary works and a number of newspaper articles. The data of one of the two samples are represented in Table 2.17.

Table 2.17: Word Length Frequencies for Croato-Serbian Text Segments (Vranić/Matković 1965)

i	f_i	p_i
1	13738	0.3420
2	12000	0.2988
3	8776	0.2185
4	4234	0.1054
5	1103	0.0275
6	253	0.0063
7	47	0.0012
8	13	0.0003
9	3	0.0001

In Table 2.17, f_i denotes the absolute and p_i the relative frequencies of i-syllable words..

Referring to the work of Fucks, and testing if their data follow a 1-displaced Poisson distribution, as suggested by Fucks, Vranić/Matković (1965b: 187) observed a clear "discrepancy from the Poisson distribution in monosyllabic and disyllabic words", at the same time seeing "indications of conformity in the distribution of three-syllable, four-syllable, and polysyllabic words." The corresponding data are represented in Figure 2.12.

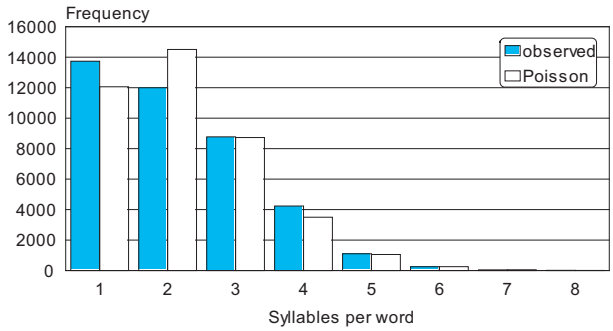


Figure 2.12: Fitting the 1-Displaced Poisson Distribution to Croato-Serbian Text Segments (Vranić/Matković 1965a,b)

We need not concentrate here on questions of the particular data structure. Rather, it is of methodological interest to see how the authors dealt with the data. Guided by the conclusion (supported by the graphical representation of Figure 2.12), the authors tested if the words of length $i \geq 3$, taken as a separate sample, follow the Poisson distribution. Calculating the corresponding χ^2 values, they reduced the whole sample of the remaining 14429 items to an artificial sample of 1000 items, retaining the proportions of the original data set. The reason for this reduction is likely to be the linear rise of χ^2 values with increasing sample size (see above, p. 23). As a result, the authors conclude "that three- and polysyllabic words in Croato-Serbian roughly follow the Poisson distribution" (ibd., 189).

In fact, a re-analysis shows that for fitting the Poisson distribution to the original sample ($N = 40167$), one obtains a rather bad discrepancy coefficient of $C = 0.0206$, whereas for that portion of words with length $i \geq 3$ one obtains $C = 0.0085$. Though convincing at first sight, the question remains why the goodness of the Poisson distribution has not been tested for that portion of words with length $i \geq 2$; curiously enough, the result is even better with $C = 0.0047$. Yet, obviously (mis-)led by the optical impression, Vranić/Matković (1965b: 194) concentrate on a modification of the first two classes, suggesting a procedure which basically implies a double modified Poisson distribution. Referring to the approaches discussed by Fucks and Bartkowiakowa/Gleich-

gewicht (see above), Vranić/Matković suggest the introduction of particular weights, which, according to their proposal, are obtained by way of the following method.

Taking the relative frequency of $p_1 = 0.342$, one obtains $\lambda = 1.079$ as that parameter of the standard (i.e., unweighted) Poisson distribution, from which $v_1 = 0.340$ results as the theoretical relative frequency:

$$v_i = \frac{\lambda^{i-1} e^{-\lambda}}{(i-1)!} \ , \quad i = 1, 2, \ldots \tag{2.31}$$

Furthermore, for $\lambda = 1.079$, one obtains $v_2 = 0.367$, and the corresponding values for the remaining frequencies $(v_3 \ldots v_n)$. Given the observation that the empirical values follow a Poisson distribution for $i \geq 3$, the authors consider it to be necessary and sufficient to represent monosyllabic and disyllabic words through superposition by way of introducing two weighting parameters a_1 and a_2 modifying the theoretical frequencies of v_1 and v_2, as obtained from (2.31), thus arriving at the weighted theoretical frequencies p'_1 and p'_2 by assuming:

$$p'_1 = a_1 \cdot v_1 \qquad\qquad p'_2 = a_1 \cdot v_2 + a_2 \cdot v_1 \ .$$

Given the condition that $p'_1 + p'_2 = p_1 + p_2 = 0.3420 + 0.2988 = 0.6408$, one has to seek the minimum for formula (2.32):

$$\begin{aligned} F(a_1, a_2) = (p'_1 - a_1 \cdot v_1)^2 + (p'_2 - a_1 \cdot v_2 - a_2 \cdot v_1) - \\ - 2\beta \cdot (v_1 + v_2 - 0.6408) \end{aligned} \tag{2.32}$$

Solving the resulting set of equations, one thus obtains the two weights $a_1 = 1.006$ and $a_2 = -0.2066$; consequently,

$$\begin{aligned} p'_1 &= 1.006 \cdot v_1 = 1.006 \cdot 0.340 = 0.342 \\ p'_2 &= 1.006 \cdot v_2 + a_2 \cdot v_1 = 1.006 \cdot 0.367 - 0.2066 \cdot 0.340 = 0.2988 \ . \end{aligned}$$

We thus obtain the weighted theoretical values NP_i of the doubly modified Poisson distribution, represented in Table 2.18.

As a re-analysis shows, the results must be regarded to be excellent, statistically confirmed by a discrepancy coefficient value of $C = 0.0030$ ($\chi^2_{df=5} = 122.18$). Still, there remain at least two major theoretical problems:

1. No interpretation is given as to why the weighting modification is necessary: is this a matter of the specific data structure, is this specific for Croatian language products?
2. Is it reasonable to stick to the Poisson distribution, though in a modified version of it, as a theoretical model, if almost two thirds of the data sample ($f_1 + f_2 \approx 64\%$) do not seem to follow it?
3. As was mentioned above, the whole sample follows a Poisson distribution not only for $i \geq 3$, but already for $i \geq 2$: consequently, in this case, only the first class would have to be modified, if it all.

Table 2.18: Fitting the Doubly Modified Poisson Distribution to Croato-Serbian Text Segments (Vranić/Matković 1965a,b)

i	f_i	NP_i
1	13738	13738.00
2	12000	12000.00
3	8776	8599.81
4	4234	4450.40
5	1103	1151.54
6	253	198.64
7	47	25.70
8	13	2.66
9	3	0.23

7. The Negative Binomial Distribution (Grotjahn 1982)

An important step in the discussion of possibly adequate distribution models for word length frequencies was Grotjahn's (1982) contribution. As can be seen above, apart from Elderton's early attempt to favor the geometric distribution, the whole discussion had focused for almost three decades on the Poisson distribution; various attempts had been undertaken to modify the Poisson distribution, due to the fact that the linguistic data under study could not be theoretically modelled by recourse to it. As the re-analyses presented in the preceding chapters have shown, neither the standard Poisson distribution nor any of its straight forward modifications can be considered to be an adequate model. Still, all the attempts discussed above, from the late 1950s until the 1980s, in one way or another, stuck to the conviction that the Poisson distribution is the one relevant model which "only" has to be modified, depending on the specific structure of linguistic data.

Grotjahn, in his attempt, opened the way for new perspectives: he not only showed that the Poisson model per se might not be an adequate model; furthermore, he initiated a discussion concentrating on the question whether one overall model could be sufficient when dealing with word length frequencies of different origin.

Taking into consideration that the 1-displaced Poisson model, basically suggested by Fucks and often, though mistakenly, called the "Fucks distribution", was still considered to be the standard model, it seems to be necessary to put some of Grotjahn's introductory remarks into the right light.

As most scholars at that time would do – and, in fact, as most scholars would do still today –, Grotjahn (1982: 46ff.), at the beginning of his ruminations, referred to the so-called "Fucks distribution". According to him, "the Fucks distribution has to be regarded a special case of a displaced Poisson distribution" (ibd., 46). As was shown above, this statement is correct only if one considers the 1-displaced Poisson distribution to be the "Fucks distribution"; in fact, however, as was shown above, the 1-displaced Fucks distribution is not more and not less than a special case of the generalized Fucks-Poisson distribution.

With this in mind, Grotjahn's own suggestions appear in a somewhat more adequate light. Given a random variable Y, representing the number of syllables per word (which may have the values $k = a, a+1, \ldots$, with $a \in \mathbb{N}_0$), we have formula (2.33) for the displaced Poisson distribution, resulting in the standard Poisson distribution for $a = 0$:

$$P(Y = k) = \frac{e^{-\lambda}\lambda^{k-a}}{(k-a)!}, \qquad k = a, a+1, ... \qquad a \in \mathbb{N}^0 \; . \tag{2.33}$$

As a starting point, Grotjahn analyzed seven letters by Goethe, written in 1782, and tested in how far the (1-displaced) Poisson distribution would prove to be an adequate model. As to the statistical testing of the findings, Grotjahn (1982: 52) suggested calculating not only χ^2 values, or their transformation into z values, but also the deviation of the empirical dispersion index (d) from its theoretical expectation (δ). As was pointed out above (cf. p. 48), the Poisson distribution can be an adequate model only in case $d \approx 1$.

However, of the concrete data analyzed by Grotjahn, only some satisfied this condition; others clearly did not, the value of d ranging from $1.01 \leq d \leq 1.32$ for the seven Goethe letters under study. Given this observation, Grotjahn arrived at two important conclusions: the first consequence is that "the displaced Poisson distribution hardly can be regarded to be an adequate model for the word length frequency distribution in German" (ibd., 55). And his second conclusion is even more important, generally stating that the Poisson model "cannot be a general law for the formation of words from syllables" (ibd., 47).

In a way, this conclusion paved the way for a new line of research. After decades of concentration on the Poisson distribution, Grotjahn was able to prove that this model alone cannot be adequate for a general theory of word length distribution. On the basis of this insight, Grotjahn further elaborated his ruminations. Replacing the Poisson parameter λ in (2.33) by $\theta - a$, and obtaining (2.34)

$$P(Y = k) = \frac{e^{-(\theta-a)}(\theta-a)^{k-a}}{(k-a)!}, \qquad k = a, a+1, ... \qquad a \in \mathbb{N}_0, \tag{2.34}$$

Grotjahn's (1982: 55) reason for this modification was as follows: a crucial implication of the Poisson distribution is the independence of individual occur-

rences. Although every single word thus may well follow a Poisson distribution, this assumption does not necessarily imply the premise that the probability is one and the same for all words; rather, it depends on factors such as (linguistic) context, theme, etc. In other words, Grotjahn further assumed that parameter θ itself is likely to be a random variable.

Now, given one follows this (reasonable) assumption, the next question is which theoretical model might be relevant for θ. Grotjahn (1982: 56ff.) tentatively assumed the gamma distribution to be adequate. Thus, the so-called negative binomial distribution (2.35) (also known as 'composed Poisson' or 'multiple Poisson' distribution) in its a-displaced form is obtained, as a result of this super-imposition of two distributions:

$$f(x;k;p) = \left(\begin{array}{c} k+x-a-1 \\ x-a \end{array} \right) p^k q^{x-a}, \qquad x = a, a+1, ... \qquad a \in \mathbb{N}_0 \tag{2.35}$$

As can be seen, for $k = 1$ and $a = 1$, one obtains the 1-displaced geometric distribution (2.2), earlier discussed by Elderton (1949) as a possible model (see above, p. 20).

$$f(x) = p \cdot q^{x-1}, \qquad x = 1, 2, \ldots \tag{2.36}$$

In fact, the negative binomial distribution had been discussed before by Brainerd (1971, 1975: 240ff.). Analyzing samples from various literary works written in English, Brainerd first tested the 1-displaced Poisson distribution and found that it "yields a poor fit in general for the works considered" (Brainerd 1975: 241). The 1-displaced Poisson distribution turned out to be an acceptable model only in the case of short passages, whereas in general, his data indicated "that a reasonable case can be made for the hypothesis that the frequencies of syllables per word follow the negative binomial distribution" (ibd., 248). In some cases, however (in fact those with $k \to 1$), also the geometric distribution (2.2) suggested by Elderton (1949) turned out to be adequate.

The negative binomial distribution does not only converge to the geometric distribution, however; under particular circumstances, it converges to the Poisson distribution: namely, if $k \to \infty, q \to 0, k \cdot q \to a$ (cf. Wimmer/Altmann 1999: 454). Therefore, as Grotjahn (1982: 71f.) rightly stated, the negative binomial distribution, too, is apt to model frequency distributions with $d \approx 1$.

With his approach, Grotjahn thus additionally succeeded in integrating earlier research, both on the geometric and the Poisson distributions, which had failed to be adequate as an overall valid model. In this context, it is of particular interest, therefore, that the negative binomial distribution is a theoretically adequate

model also for data with $d > 1$. Given the theoretical values for σ^2 and μ

$$\sigma^2 = \frac{k \cdot q^2}{p^2} + \frac{k \cdot p}{p}$$
$$\mu = \frac{k \cdot q}{p} ,$$

it can easily be shown that for the negative binomial distribution,

$$\delta = 1 + \frac{1-p}{p} > 1 . \tag{2.37}$$

As Grotjahn (1982: 61) concludes, the negative binomial distribution therefore should be taken into account for empirical distributions with $d > 1$. A comparison of German corpus data from Meier's (1967) *Deutsche Sprachstatistik* clearly proves Grotjahn's argument to be reasonable. The data are reproduced in Table 2.19, which contains the theoretical values both for the Poisson and the negative binomial distributions. In addition to the χ^2 values, given by Grotjahn, Table 2.19 also contains the values of the discrepancy coefficient C discussed above (cf. p. 23), which are calculated anew, here.

Table 2.19: Fitting the Negative Binomial and Poisson Distributions to German Data from Meier's Corpus (Grotjahn 1982)

x	f_x	neg. binom. d. NP_x	Poisson d. NP_x
1	25940	25827.1	22357.1
2	14113	14174.9	17994.8
3	5567	6144.5	7241.8
4	2973	2427.2	1942.9
5	1057	912.1	391.0
6	264	332.2	62.9
7	74	118.5	8.4
8	10	41.6	1.0
≥ 9	2	21.9	0.1
$N = 50000$		$\chi^2 = 273.72$	$\chi^2 = 4752.17$
		$C = 0.005$	$C = 0.095$

As can be seen, the negative binomial distribution yields significantly better results as compared to the Poisson model. The results are graphically represented in Figure 2.13.

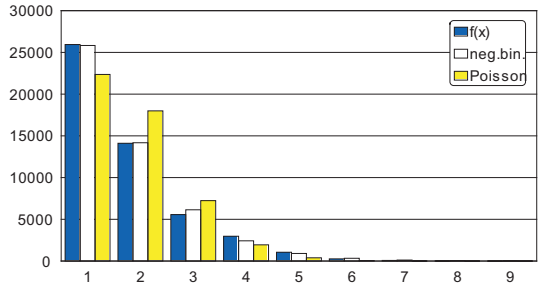


Figure 2.13: Observed and Expected Word Length Frequencies for Meier's German Corpus (Grotjahn 1982)

Concluding, it seems important to emphasize that Grotjahn's (1982: 74) overall advice was that the negative binomial distribution should be taken into account as *one* possible model for word length frequencies, not as *the* only general model. Still, it is tempting to see in how far the negative binomial distribution is able to model the data of nine languages, given by Fucks (cf. Table 2.9, p. 41). Table 2.20 represents the corresponding results, including the estimated values for the parameters k and p.

Table 2.20: Fitting the Negative Binomial Distribution to Fucks' Data From Nine Languages

	English	**German**	**Esperanto**	**Arabic**	**Greek**
$\hat{k}$	1.04	3.62	597.59	9.89	5.09
$\hat{p}$	0.72	0.85	0.99	0.90	0.82
C	0.0026	0.0019	0.0026	0.1503	0.0078

	Japanese	**Russian**	**Latin**	**Turkish**
$\hat{k}$	4.79	7.71	12.47	13.11
$\hat{p}$	0.81	0.86	0.90	0.90
C	0.0036	0.0078	0.0330	0.0440

From Table 2.20, two things can be nicely seen:

1. For Esperanto – the only 'language' with a really convincing fitting result of the Poisson distribution (cf. Table 2.10, p. 43) – both parameters behave as predicted: $k \to \infty$, and $q = (1 - p) \to 0$.
2. Particularly from the results for Arabian, Latin, and Turkish (all with $C > 0.02$), it is evident that the negative binomial distribution indeed cannot be an overall adequate model.

In so far, historically speaking, Grotjahn's (1982: 73) final conclusion that for German texts, the negative binomial distribution leads to better results almost without exception, is not as important as the general insight of his study: namely, that instead of looking for one general model one should rather try to concentrate on a variety of distributions which are able to represent a valid "law of word formation from syllables".

8. The Poisson-Uniform Distribution: Kromer (2001/02)

Based on Grotjahn's (1982) observation as to frequent discrepancies between empirical data and theoretical models thereof, Grotjahn/Altmann (1993) generalized the importance of this finding by methodologically reflecting principal problems of word length studies. Their discussion is of unchanged importance, still today, since many more recent studies in this field do not seem to pay sufficient attention to the ideas expressed almost a decade ago.

Before discussing these important reflections, one more model should be discussed, however, to which attention has recently been directed by Kromer (2001a,b,c; 2002). In this case, we are concerned with the Poisson-uniform distribution, also called Poisson-rectangular distribution (cf. Wimmer/Altmann 1999: 524f.). Whereas Grotjahn's combination of the Poisson distribution with a second model (i.e., the gamma distribution), resulted in a specific distribution in its own right (namely, the negative binomial distribution), this is not the case with Kromer's combination of the Poisson distribution (2.8) with the uniform (rectangular) distribution:

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b\ . \tag{2.38}$$

As a result of combining the rectangular distribution (2.38) with the Poisson distribution (2.8), one obtains the Poisson uniform distribution:

$$P_x = (b-a)^{-1}\left[e^{-a}\sum_{j=0}^{x}\frac{a^j}{j!} - e^{-b}\sum_{j=0}^{x}\frac{b^j}{j!}\right], \quad x = 0, 1, 2, \ldots \tag{2.39}$$

Here, a necessary condition is that $b > a \geq 0$. In his approach, Kromer (2001a) derived the Poisson-uniform distribution along a different theoretical way, which need not be discussed here in detail. With regard to formula (2.39), this results in a replacement of parameters a and b by $(\lambda_1 - 1)$ and (λ_2-1), thus leading to the following 1-displaced form (with the support $x = 1, 2, 3, \ldots$):

$$P_x = \frac{1}{\lambda_2 - \lambda_1} \left[e^{-(\lambda_1 - 1)} \sum_{j=1}^{x} \frac{(\lambda_1 - 1)^{j-1}}{(j-1)!} - e^{-(\lambda_2 - 1)} \sum_{j=1}^{x} \frac{(\lambda_2 - 1)^{j-1}}{(j-1)!} \right]. \tag{2.39a}$$

Kromer then defined the mean of the distribution to be

$$\lambda_0 = \frac{\lambda_1 + \lambda_2}{2}. \tag{2.40}$$

A simple transformation of this equation leads to $\lambda_2 = 2 \cdot \lambda_0 - \lambda_1$. As a result, one thus obtains λ_2 as depending on λ_1 which remains to be estimated. With regard to this question, Kromer (2001a: 95) discusses two methods: the method of moments, and the method of χ^2 minimization.

Since, as a result, Kromer does not favor the method of moments, he unfortunately does not deem the system of equations necessary to arrive at a solution for λ_1. It would be too much, here, to completely derive the two relevant equations anew. It may suffice therefore to say that the first equation can easily be derived from (2.40); as to the second necessary equation, Kromer (2001a: 95) refers to the second initial moments of the empirical (m_2') and the theoretical (μ_2') distributions (cf. page 52):

$$m_2' = \frac{1}{N} \sum_x x^2 \cdot f_x \qquad\qquad \mu_2' = \sum_x x^2 \cdot P_x$$

One thus obtains the following system of equations:

(a)	$0 = \lambda_1 + \lambda_2 - 2\bar{x}$
(b)	$0 = 6m_2' - 2\lambda_1^2 - 3\lambda_1 - 2\lambda_2^2 - 3\lambda_2 - 2\lambda_1\lambda_2 + 6$

In empirically testing the appropriateness of his model, Kromer (2001a) used data from Best's (1997) study on German-language journalistic texts from an Austrian journal. Best, in turn, had argued in favor of the negative binomial distribution discussed above, as an adequate model.

The results obtained for these data need not be presented here, since they can easily be taken from the table given by Kromer (2001a: 93). It is more important to state that Kromer (2001a: 95), as a result of his analyses, found "that the method of moments leads to an unsatisfactory approximation of the

empirical distribution by the theoretical one owing to the strong dependence of the second moment of the distribution on random factors". Kromer therefore suggested not to use this procedure, and to prefer the method of χ^2 minimization.

In the case of this method, we are concerned with a merely numerical solution, fitting λ_1 by minimizing the χ^2 value. Instead of presenting the results of Kromer's fittings, it might be tempting to re-analyze once again Fucks' data (cf. Table 2.9). These data have been repeatedly analyzed above, among others with regard to the negative binomial distribution (cf. Table 2.20, p. 65). Since the negative binomial distribution had proven not to be an adequate model for Latin, Arabic, and Turkish, it is interesting to see the results one obtains with Kromer's model.

Table 2.21 presents the corresponding results. In addition to the values for $\hat{\lambda}_1$ and $\hat{\lambda}_2$, obtained according to the two methods described above, Table 2.21 also contains the results one obtains for the 1-displaced Poisson-uniform distribution, using iterative methods incorporating relevant special software (ALTMANN-Fitter, version 2.1, 2000).

It can clearly be seen that for the 1-displaced Poisson-uniform distribution (with $b > a \geq 0$), there are solutions for all data sets, although for four of the nine languages, the results cannot be called satisfying ($C > 0.02$): these four languages are English, Arabic, Latin, and Turkish. As compared to this, the results for Kromer's modification are better in all cases. Additionally, they prove to be interesting in a different aspect, depending on the manner of estimating λ_1 (and, consequently, of λ_2). Using the method of moments, it turns out that in four of the nine cases (Esperanto, Arabic, Latin, and Turkish), no acceptable solutions are obtained. However, for these four cases, too, acceptable results are obtained with the χ^2 minimization method. Interestingly the values for λ_1 and λ_2, obtained with this method, are almost identical, differing only after the fifth or higher decimal (thus, $\lambda_1 \approx \lambda_2 \approx \lambda_0$).

Now, what is the reason for no satisfying results being obtained, according to the method of moments? Let us once again try to explain this referring to the dispersion quotient δ discussed above (cf. p. 47). As can be seen above, $\delta = Var(X)/[E(X)-1]$. Now, given that, for Kromer's version of the Poisson-uniform distribution in its 1-displaced form, we have the theoretical first and second moments:

$$\begin{aligned}
\mu_1 =& \frac{(\lambda_1 - 1) + (\lambda_2 - 1)}{2} + 1 = \frac{\lambda_1 + \lambda_2}{2} \\
\mu_2 =& \frac{[(\lambda_1 - 1) - (\lambda_2 - 1)]^2}{12} + \frac{(\lambda_1 - 1) + (\lambda_2 - 1)}{2} \\
=& \frac{(\lambda_1 - \lambda_2)^2}{12} + \frac{\lambda_1 + \lambda_2 - 2}{2}
\end{aligned}$$

Table 2.21: Fitting the 1-Displaced Poisson-Uniform Distribution to Fucks' Data From Nine Languages

	English	**German**	**Esperanto**	**Arabic**	**Greek**
$b > a \geq 0$					
$\hat{a}$	0.0497	0.1497	0.4675	0.6101	0.3197
$\hat{b}$	0.8148	1.1235	1.3432	1.6686	1.9199
C	0.0288	0.0029	0.0068	0.1409	0.0065
$\bar{x}, m_2'$					
$\hat{\lambda}_1$	0.7178	1.0567	∅	∅	1.2587
$\hat{\lambda}_2$	2.0950	2.2100	∅	∅	2.9625
C	0.0028	0.0027	–	–	0.0047
χ^2-min.					
$\hat{\lambda}_1$	0.7528	1.0904	1.8971	2.1032	1.1556
$\hat{\lambda}_2$	2.0600	2.1763	1.8971	2.1032	3.0656
C	0.0021	0.0024	0.0023	0.1071	0.0023
$d > 1$	✓	✓	–	–	✓

	Japanese	**Russian**	**Latin**	**Turkish**
$b > a \geq 0$				
$\hat{a}$	0.3457	0.3720	0.8373	0.8635
$\hat{b}$	1.9401	2,0900	1.9942	2.0859
C	0.0054	0.0054	0.0282	0.0391
$\bar{x}, m_2'$				
$\hat{\lambda}_1$	1.2451	1.4619	∅	∅
$\hat{\lambda}_2$	3.0199	2.9918	∅	∅
C	0.0053	0.0060	–	–
χ^2-min.				
$\hat{\lambda}_1$	1.3122	1.3088	2.3894	2.4588
$\hat{\lambda}_2$	2.9528	3.1449	2.3894	2.4588
C	0.0037	0.0037	0.0166	0.0207
$d > 1$	✓	✓	–	–

As to the theoretical dispersion quotient δ, we thus obtain the following equation:

$$\begin{aligned}\delta = \frac{Var(X)}{E(X)-1} &= \frac{\frac{(\lambda_1-\lambda_2)^2}{12}+\frac{\lambda_1+\lambda_2-2}{2}}{\frac{\lambda_1+\lambda_2}{2}-1} \\ &= \frac{\frac{(\lambda_1-\lambda_2)^2+6\lambda_1+6\lambda_2-12}{12}}{\frac{\lambda_1+\lambda_2-2}{2}} = \frac{(\lambda_1-\lambda_2)^2+6(\lambda_1+\lambda_2-2)}{6(\lambda_1+\lambda_2-2)} \\ &= \frac{(\lambda_1-\lambda_2)^2}{6\,(\lambda_1+\lambda_2-2)}+1\,.\end{aligned}$$

Because $(\lambda_1-\lambda_2)^2$ is positive, and because $\lambda_1 > 1$ and $\lambda_2 > 1$ by definition, $(\lambda_1+\lambda_2-2)$ must be positive, as well; therefore, the quotient

$$Q_\delta = \frac{(\lambda_1-\lambda_2)^2}{6\,(\lambda_1+\lambda_2-2)} > 0 \tag{2.41}$$

must be positive as well. Consequently, for the 1-displaced Poisson-uniform distribution to be fitted with the method of moments, a necessary condition is that the dispersion quotient is $d > 1$. Empirically, this is proven by the results represented in Table 2.21: here, for those cases with $d \leq 1$, fitting Kromer's modification of the Poisson-uniform distribution with the method of moments fails.

Additionally, this circumstance explains why in these cases, we have almost identical values for λ_1 and λ_2 (i.e., $\lambda_1 \approx \lambda_2$): As can be shown, the dispersion quotient of the 1-displaced Poisson-uniform distribution is $\delta = 1$, only in the case that the quotient $Q_\delta = 0$ – cf. equation (2.41), as to this point. This however, is the case only if $\lambda_1 = \lambda_2$. Actually, this explains Kromer's assumption that for $\lambda_1 = \lambda_2$, the 1-displaced Poisson-uniform "degenerates" to the 1-displaced Poisson distribution, where, by definition, $\delta = 1$.[13]

According to Kromer (2001a: 96, 2001b: 74), the model proposed by him "degenerates" into the Poisson (Čebanov-Fucks) distribution with $\lambda_1 = \lambda_0$ (and correspondingly $\lambda_2 = \lambda_0$). In principle, this assumption is correct; strictly speaking, however, it would be more correct to say that for $\lambda_1 \cong \lambda_2$, the 1-displaced Poisson-uniform distribution can be approximated by the Poisson distribution. For the sake of clarity, the approximation of the 1-displaced

[13] From this perspective, it is no wonder that the C values obtained for the Poisson-uniform distribution by way of the χ^2 minimization method are almost the same, or even identical to those obtained for the Poisson distribution (cf. Table 2.10, p. 43).

Poisson-uniform distribution suggested by Kromer (personal communication) shall be demonstrated here; it is relevant for those cases when parameter a converges with parameter b in equation (2.39). In these cases, when $b = a + \varepsilon$ with $\varepsilon \to 0$, we first replace b with $a + \varepsilon$ in equation (2.39), thus obtaining formula (2.39'):

$$P_x = \frac{1}{\varepsilon}\left(e^{-a}\sum_{j=0}^{x}\frac{a^j}{j!} - e^{-a-\varepsilon}\sum_{j=0}^{x}\frac{(a+\varepsilon)^j}{j!}\right) . \qquad (2.39')$$

In the next step, the binomial expression $(a + \varepsilon)^j$ from equation (2.39') is replaced with its first two terms, i.e.,

$$(a+\varepsilon)^j = \left(a\left(1+\frac{\varepsilon}{a}\right)\right)^j = a^j\left(1+\frac{\varepsilon}{a}\right)^j = a^j\left(1+\frac{\varepsilon}{a}\cdot j+\ldots\right) \approx$$
$$\approx a^j\left(1+\varepsilon\cdot j\cdot a^{-1}\right) = a^j\left(1+\frac{\varepsilon\cdot j}{a}\right) = a^j + a^{j-1}\varepsilon\cdot j,$$

thus obtaining (2.39")

$$P_x = \frac{1}{\varepsilon}\left\{e^{-a}\sum_{j=0}^{x}\frac{a^j}{j!} - e^{-a}\cdot e^{-\varepsilon}\left(\sum_{j=0}^{x}\frac{a^j}{j!} + \sum_{j=0}^{x}\frac{a^{j-1}\varepsilon j}{j!}\right)\right\} . \qquad (2.39")$$

Finally, function $e^{-\varepsilon}$ in equation (2.39") is approximated by the first two terms of the Taylor series of this function, resulting in $1 - \varepsilon$, thus stepwise receiving the ordinary Poisson distribution:

$$P_x = e^{-a}\left(\sum_{j=0}^{x}\frac{a^j}{j!} - \sum_{j=0}^{x}\frac{j\cdot a^{j-1}}{j!}\right) = e^{-a}\frac{a^x}{x!} . \qquad (2.39"')$$

Yet, we are concerned here with an approximation of the Poisson-uniform distribution, not with its convergence to the Poisson distribution, since $\lambda_1 = \lambda_2$ would result in zero for the first part of equation (2.39a), and the second part of (2.39a) would make no sense either, also resulting in 0 (cf. p. 67).

Anyway, Kromer's (2001c) further observation – based on the results obtained by the χ^2 minimization – saying that there seems to be a direct dependence of λ_1 on λ_0, is of utmost importance and deserves further attention. In fact, in addition to his assumption that this is the case for homogeneous texts of a given genre only, a re-analysis of Fucks' data (cf. p. 41) as to this question corroborates and extends Kromer's findings; although these data are based on mixed corpora of the languages under study, there is a clear linear dependence of λ_1 on λ_0, for these data as well ($R^2 = 0.91$).

In this respect, another assumption of Kromer's might turn out to be important, here. This assumption is as plausible and as far-reaching, since Kromer postulates two invariant parameters (I and α, in his terminology) to be at work in the generation of word length frequencies. According to Kromer, the first of these two parameters (I) is supposed to be an invariant parameter for the given language, being defined as $I = (\lambda_0 - 1) \cdot (\lambda_1 - \lambda_{1min})$. It is important to note that parameter λ_{1min} should not be confounded here with the result of the χ^2 minimization described above; rather, it is the lower limit of λ_1. On the basis of his analyses, Kromer (2001b,c, 2002) assumes λ_{1min} to be 0.5, approximately. The second parameter α can be derived from the equation $\lambda_1 = \alpha \cdot \lambda_{1min} + (1 - \alpha) \cdot \lambda_0$. Consequently, it is defined as $\alpha = (\lambda_0 - \lambda_1)/(\lambda_0 - \lambda_{1min})$.

According to Kromer, both parameters (I and α) allow for a direct linguistic interpretation. Parameter I, according to him, expresses something like the specifics of a given language (i.e., the degree of a language's syntheticity (Kromer 2001c). As opposed to this, parameter α characterizes the degree of completion of synergetic processes optimizing the code of the given language. According to Kromer (2001c), $\alpha \in (0, 1)$ for real texts, with $\alpha \approx 0.3 - 0.6$ for simple genres (such as letters or children's literature), and $\alpha \approx 0.8$ for more complex genres (such as journalistic or scientific texts).

Unfortunately, most of the above-mentioned papers (Kromer 2001b,c; 2002) have the status of abstracts, rather than of complete papers; as a consequence, only scarce empirical data are presented which might prove the claims brought forth on a broader empirical basis. In summary, one can thus first of all say that Kromer's modification of the Poisson-uniform distribution, as well as the original Poisson-uniform distribution, turns out to be a model which has thus proven its adequacy for linguistic material from various languages. Particularly Kromer's further hope to find language- and text-specific invariants deserves further study. If his assumption should bear closer examination on a broader empirical basis, this might as well explain why we are concerned here with a mixture of two distributions. However, one must ask the question, why it is only the rectangular distribution which comes into play here, as one of two components. In other words: Wouldn't it be more reasonable to look for a model which by way of additional parameters, or by way of parameters taking extreme values (such as $0, 1$, or ∞) allows for transitions between different distribution models, some of them being special cases, or generalizations, of some superordinate model? Strangely enough, it is just the Poisson-uniform distribution, which converges to almost no other distribution, not even to the Poisson distribution, as can be seen above (for details, cf. Wimmer/Altmann 1999: 524).

Ultimately, this observation leads us back to the conclusion drawn at the end of the last chapter, when the necessity to discuss the problems of word length

studies from a methodological point of view was mentioned. This discussion was initiated by Grotjahn and Altmann as early as in 1993, and it seems important to call to mind the most important arguments brought forth some ten years ago.

9. Theoretical and Methodological Reflections: Grotjahn/Altmann (1993)

This is not to say that no attention has been paid to the individual points raised by Grotjahn and Altmann. Yet, only recently systematic studies have been undertaken to solve just the methodological problems by way of empirical studies. It would lead too far, and in fact be redundant, to repeat the authors' central arguments here. Nevertheless, most of the ideas discussed – Grotjahn and Altmann combined them in six groups of practical and theoretical problems – are of unchanged importance for contemporary word length studies, which makes it reasonable to summarize at least the most important points, and comment on them from a contemporary point of view.

a. *The problem of the unit of measurement.*– As to this question, it turns out to be of importance what Ferdinand de Saussure stated about a century ago, namely, that there are no positive facts in language. In other words: There can be no a priori decision as to what a word is, or in what units word length can be measured. Meanwhile, in contemporary theories of science, linguistics is no exception to the rule: there is hardly any science which would not acknowledge, to one degree or another, that it has to define its object, first, and that constructive processes are at work in doing so. The relevant thing here is that measuring is (made) possible, as an important thing in the construction of theory. As Grotjahn/Altmann (1993: 142) state with regard to word length, there are three basic types of measurement which can be distinguished: graphic (e.g. letters), phonetic (sounds, phonemes, syllables, etc.), and semantic (morphemes). And, as a consequence, it is obvious "that the choice of the unit of measurement strongly effects the model of word length to be constructed" (ibd., 143).
   What has not yet been studied is whether there are particular dependencies between the results obtained on the basis of different measurement units; it goes without saying that, if they exist, they are highly likely to be language-specific.
   Also, it should be noted that this problem does not only concern the unit of measurement, but also the object under study: the word. It is not even the problem of compound words, abbreviation and acronyms, or numbers and digits, which comes into play here, or the distinction between word forms and lexemes (lemmas) – rather it is the decision whether a word is to be defined on a graphemic, orthographic-graphemic, or phonological level.

Defining not only the measurement unit, but the unit under investigation itself, we are thus faced with the very same problems, only on a different (linguistic, or meta-linguistic) level.

b. *The population problem.*– As Grotjahn/Altmann (1993: 143ff.) rightly state, the result can be expected to be different, depending on whether the material under study is taken from a dictionary, from a frequency dictionary, or from texts. On the one hand, when one is concerned with "ordinary" dictionaries, one has to be aware of the fact that attention is paid neither to frequency nor to the frequency of particular word forms; on the other hand, in the case of frequency dictionaries, the question is what kind of linguistics material has been used to establish the frequencies. And, as far as a text is considered to be the basic unit of study, one must ask what a 'text' is: is it a chapter of a novel, or a book composed of several chapters, or the complete novel?
Again, as to these questions, there are hardly any systematic studies which would aim at a comparison of results obtained on an empirical basis. More often than not, letters as a specific text type have been considered to be "prototypical" texts, optimally representing language due to the interweaving of oral and written components. However, there are some dozens of different types of letters, which can be proven to follow different rules, and which even more clearly differ from other text types.
One is thus concerned, in one way or another, with the problem of data homogeneity: therefore, one should not only keep apart dictionaries (of various kinds) on the one hand, and texts, on the other – rather, one should also make clear distinctions between complete 'texts', text segments (i.e., randomly chosen parts of texts), text mixtures (i.e., combinations of texts, from the combination of two texts up to the level of complete corpora), and text cumulations (i.e., that type of text, which is deliberately composed of subsequent units).

c. *The goodness-of-fit problem.*– Whereas Grotjahn/Altmann (1993: 147ff.) present an extensive discussion of this problem, it has become usual, by now, to base any kind of test on Pearson's χ^2 test. And, since it is well-known that differences are more likely to be significant for large samples (since the χ^2 value increases linearly with sample sizes), it has become the norm to calculating the discrepancy coefficient $C = \chi^2/N$, with two conventional deviation boundaries: $0.01 < C < 0.02$ ("good fit"), and $C < 0.01$ ("very good fit"). The crucial unsolved question, in this field, is not so much if these boundaries are reasonable – in fact, there are some studies which use the $C < 0.05$ boundary, otherwise not obtaining acceptable results. Rather, the question is, what is a small text, and where does a large text start? And why do we, in some cases, obtain significant C values when $p(\chi^2)$ is significant, too, but in other cases do not?

d. The *problem of the interrelationship of linguistic properties.*– Under this heading, Grotjahn/Altmann (1993: 150) analyzed a number of linguistic properties interrelated with word length. What they have in mind are intralinguistic factors which concern the synergetic organization of language, and thus the interrelationship between word length factors such as size of the dictionary, or the phoneme inventory of the given language, word frequency, or sentence length in a given text (to name but a few examples).
The factors enumerated by Grotjahn/Altmann all contribute to what may be called the boundary conditions of the scientific study of *language*. As soon as the interest shifts from language, as a more or less abstract system, to the object of some (real, fictitious, imagined, or even virtual) communicative act, between some producer and some recipient, we are not concerned with language, any more, but with *text*. Consequently, there are more factors to be taken into account forming the boundary conditions, factors such as author-specific, or genre-dependent conditions. Ultimately, we are on the borderline here, between quantitative linguistics and quantitative text analysis, and the additional factors are, indeed, more language-related than intralinguistic in the strict sense of the word. However, these factors cannot be ignored, as soon as running texts are taken as material; it might be useful, therefore, to extend the problem area outlined by Grotjahn/Altmann and term it *the problem of language-related and text-related influence factors*. It should be mentioned, however, that very little is known about such factors, and systematic work on this problem has only just begun.

e. *The modelling problem.*– As Grotjahn/Altmann (1993: 146f.) state, it is very unlikely that one single model should be sufficient for the various types of data involved. Rather one would, as they claim, "expect one specific model for each data type" (ibd., 146). Grotjahn/Altmann mainly had in mind the distinctions of different populations, as they were discussed above (i.e. dictionary vs. frequency dictionary, vs. text, etc.); the expectation brought forth by them, however, ultimately results in the possibility that there might be single models for specific boundary conditions (i.e. for specific languages, for texts of a given author written in a particular language, or for specific text types in a given language, etc.).
The options discussed by Grotjahn/Altmann (1993) are relevant, still today, and they can be categorized as follows: (i) find a single model for the data under study; (ii) find a compound model, a convolution, or a mixture of models, for the data under study. As can be seen, the aim may be different with regard to the particular research object, and it may change from case to case; what is of crucial relevance, then, is rather the question of interpretability and explanation of data and their theoretical modelling.

f. *The problem of explanation.*– As Grotjahn/Altmann (1993: 150f.) correctly state, striving for explanation is the primary and ultimate aim of science. Consequently, in order to obtain an explanation of the nature of word length, one must discover the mechanism generating it, hereby taking into account the necessary boundary conditions. Thus far, we cannot directly concentrate on the study of particular boundary conditions, since we do not know enough about the general system mechanism at work. Consequently, contemporary research involves three different kinds of orientation: first, we have many bottom-up oriented, partly in the form of ad-hoc solutions for particular problems, partly in the form of inductive research; second, we have top-down oriented, deductive research, aiming at the formulation of general laws and models; and finally, we have much exploratory work, which may be called abductive by nature, since it is characterized by constant hypothesis testing, possibly resulting in a modification of higher-level hypotheses.
As to a possible way of achieving these goals, Grotjahn/Altmann (1993: 147) have suggested to pursue the "synergetic" approach of modelling. In this framework, it is not necessary to know the probabilities of all individual frequency classes; rather, it is sufficient to know the (relative) difference between two neighboring classes, e.g.

$$D = \frac{P_x - P_{x-1}}{P_{x-1}} \text{ , or } D = P_x - P_{x-1}$$

and set up theories about D. Ultimately, this line of research has in fact provided the most important research impulses in the 1990s, which shall be discussed in detail below.

10. From the Synergetic Approach to a Unified Theory of Linguistic Laws (Altmann/Grotjahn/Köhler/Wimmer)

In their 1994 contribution "Towards a theory of word length distribution", Wimmer et al. regarded word length as a "part of several control cycles which maintain the self-organization in language" (ibd., 101). Generally assuming that the distribution of word length in the lexicon and in texts follows a law, the authors further claim that the "empirical distributions actually observed can be represented as specifications of this law according to the boundary and subsidiary conditions which they are subject to in spite of the all-pervasive creativity of speakers/writers" (ibd., 101).

In their search for relevant regularities in the organization of word length, Wimmer et al. (1994: 101) then assume that the various word length classes do not evolve independently of each other, thus obtaining the following basic mechanism:

$$P_x = g(x)P_{x-1} \; . \tag{2.42}$$

With regard to previous results from synergetic linguistics, particular research on the so-called "Menzerath law", modelling the regulation of the size of (sub)systems by the size of the corresponding supersystems, Wimmer et al. state that in elementary cases the function $g(x)$ in (2.42) has the form

$$g(x) = ax^{-b} \; . \tag{2.43}$$

Based on these assumptions, Wimmer et al. (1994: 101ff.) distinguish three levels, if one wants, as to the synergetic modelling of word length distribution: (a) elementary form, (b) modification, and (c) complication.

(a) The most elementary, basic organization of a word length distribution would follow the difference equation

$$P_{x+1} = \frac{a}{(x+1)^b} P_x \; , \quad x = 0, 1, 2, \ldots \quad a, b > 0 \; . \tag{2.44}$$

Depending on whether there are 0-syllable words or not (i.e., $P_0 = 0$ or $P_0 \neq 0$), one obtains one of the two following formulas (2.45) or (2.45a), which are identical except for translation, i.e. either:

$$P_x = \frac{a^x}{(x!)^b} P_0 \; , \quad x = 0, 1, 2, \ldots \quad a, b > 0 \tag{2.45}$$

or, in 1-displaced form:

$$P_x = \frac{a^{x-1}}{(x-1)!^b} P_1 \; , \quad x = 1, 2, 3, \ldots \quad a, b > 0 \; , \tag{2.45a}$$

This finally results in the so-called Conway-Maxwell-Poisson distribution (cf. Wimmer et al. 1994: 102; Wimmer/Altmann 1999: 103), i.e.:

$$P_x = \frac{a^x}{(x!)^b \, T_0}, \quad x = 0, 1, 2, \ldots, \quad a \geq 0, b > 0, \quad T_0 = \sum_{j=0}^{\infty} \frac{a^j}{(j!)^b} \tag{2.46}$$

with T_0 as norming constant. This model was already discussed above, in its 1-displaced form (2.7), when discussing the Merkytė geometric distribution (cf. p. 26). It has also been found to be an adequate model for word length frequencies from a Slovenian frequency dictionary (Grzybek 2001).

(b) As to the second level of modelling ("first order extensions"), Wimmer et al. (1994: 102) suggested to set parameter $b = 1$ in equation (2.43) and to modify the proportionality function. After corresponding re-parametrizations,

these modifications result in well-known distribution models. In 1994, Wimmer et al. wrote that $g(x)$-functions like the following had been found:

$$\begin{aligned} \text{Hyper-Poisson:} \quad & g(x) = \frac{a}{(c+x)} \\ \text{Hyper-Pascal:} \quad & g(x) = \frac{(a+bx)}{(c+dx)} \\ \text{negative binomial:} \quad & g(x) = \frac{(a+bx)}{cx} \ . \end{aligned}$$

This system of modifications was further elaborated by Wimmer/Altmann in 1996, and shall be presented in detail, below (cf. p. 81ff.).

(c) The third level of modelling is more complex: as Wimmer et al. (1994: 102f.) say, in these more complex models "it is not appropriate to take into account only the neighboring class $(x-1)$. The set of word length classes is organized as a whole, i.e., the class of length x is proportional to all other classes of smaller length $j(j = 1, 2, \ldots, x)$." This can be written as

$$P_x = \sum_{j=1}^{x} h(j) P_{x-j} \ .$$

Inserting the original proportionality function $g(x)$ thus yields (2.47), rendering (2.42) a special case of this more complex form:

$$P_x = g(x) \sum_{j=1}^{x} h(j) P_{x-j} \ . \tag{2.47}$$

If one again chooses $g(x) = a \cdot x^{-b}$ with $b = 1$, as in the case of the first order extensions (b), this results in $g(x) = a/x$; if one furthermore defines $h(j) = j \prod_j$ – where $\prod_j$ itself is a probability function of a variable J –, then the probability P_x fulfills the necessary conditions $P_x \geq 0$ and $\sum_x P_x = 1$.

Now, different distributions may be inserted for $\prod_j$. Thus, inserting the Borel distribution (cf. Wimmer/Altmann 1999: 50f.)

$$P_x = \frac{e^{-ax} \cdot x^{x-1} \cdot x^{x-2}}{(x-1)!} \ , \qquad x = 1, 2, 3, \ldots \quad 0 \leq a < 1 \tag{2.48}$$

for $\prod_j$ in $h(j) = j \prod_j$, yields

$$P_x = \frac{a}{x} \sum_{j=1}^{x} \frac{e^{-bj} \, (bj)^{j-1}}{(j-1)!} P_{x-j} \ . \tag{2.49}$$

The solution of this is a specific generalized Poisson distribution (GPD), usually called Consul-Jain-Poisson distribution (cf. Wimmer/Altmann 1999: 93ff.):

$$\begin{aligned} P_0 &= e^{-a}\,, \\ P_x &= \frac{a\,(a+bx)^{x-1}\,e^{-(a+bx)}}{x!}\,, \quad x = 1, 2, 3, \ldots \end{aligned} \tag{2.50}$$

It can easily be seen that for $b = 0$, the standard Poisson is a special case of the GPD. The parameters a and b of the GPD are independent of each other; there are a number of theoretical restrictions for them, which need not be discussed here in detail (cf. Antić/Grzybek/Stadlober 2005a,b). Irrespective of these restrictions, already Wimmer et al. (1994: 103) stated that the application of the GPD has turned out to be especially promising, and, by way of an example, they referred to the results of fitting the generalized Poisson distribution to the data of a Turkish poem. These observations are supported by recent studies in which Stadlober (2003) analyzed this distribution in detail and tested its adequacy for linguistic data. Comparing the GPD with Fucks' generalization of the Poisson distribution (and its special cases), Stadlober demonstrated that the GPD is extremely flexible, and therefore able to model heterogeneous linguistic data. The flexibility is due to specific properties of the mean and the variance of the GPD, which, in its one-displaced form, are:

$$\begin{aligned} \mu &= E(X) = \frac{a+1-b}{1-b} \quad \text{and} \\ \sigma^2 &= Var(X) = \frac{a}{(1-b)^3}\,. \end{aligned}$$

Given these characteristics, we may easily compute δ, as was done in the case of the generalized Fucks distribution and its special cases (see above):

$$\delta = \frac{Var(X)}{E(X)-1} = \frac{1}{(1-b)^2} \geq \frac{1}{4}\,.$$

Thus, whereas the Poisson distribution turned out to be an adequate model for empirical distributions with $d \approx 1$, the 2-parameter Dacey-Poisson distribution with $d < 1$, and the 3-parameter Fucks distribution with $d \geq 0.75$, the GPD proves to be an alternative model for empirical distributions with $D \geq 0.25$ (cf. Stadlober 2004). It is interesting to see, therefore, in how far the GPD is able to model Fucks' data from nine languages, represented in Table 2.9, repeatedly analyzed above; the results taken from Stadlober (2003) are given in Table 2.22.

As can be seen, the results are good or even excellent in all cases; in fact, as opposed to all other distributions discussed above, the Consul-Jain GPD is able to model all data samples given by Fucks. It can also be seen from Table 2.22

Table 2.22: Fitting the Generalized Poisson Distribution (*GPD*) to Fucks' Data From Nine Languages

	English	**German**	**Esperanto**	**Arabic**	**Greek**
$\hat{a}$	0.3448	0.5842	0.9198	1.4285	1.0063
$\hat{b}$	0.1515	0.0775	−0.0254	−0.2949	0.0939
C	0.0030	0.0019	0.0014	0.0121	0.0072

	Japanese	**Russian**	**Latin**	**Turkish**
$\hat{a}$	1.0204	1.1395	1.4892	1.6295
$\hat{b}$	0.0990	0.0712	0.0719	−0.1170
C	0.0037	0.0078	0.0092	0.0053

that the empirical findings confirm the theoretical assumption that there is no dependence between the parameters a and b – this makes it rather unlikely that it might be possible to arrive at a direct interpretation of the results. In this respect, i.e. as to an interpretation of the results, an even more important question remains to be answered, already raised by Wimmer et al. (1994: 103), namely what might be a linguistic justification for the use of the Borel distribution.

As to this problem, it seems however important to state that this is not a problem specifically related to the GPD; rather, any mixture of distributions will cause the very same problems. From this perspective, the crucial question as to possible interpretation remains open for Fucks' generalization too, however, as well as for any other distribution implying weights, as long as no reason can be given for the amount of the specific weights of the elements in the ε-spectrum.

In this respect, it is important that other distributions which imply no mixtures can also be derived from (2.47). Thus, as Wimmer/Altmann (1996: 126ff.) have shown in detail, the probability generating function of X in (2.47) is

$$G(t) = e^{a[H(t)-1]}, \tag{2.51}$$

which leads to the so-called generalized Poisson distributions; the specific solution merely depends on the choice of $H(t)$. Now, if one sets, for example, $H(t) = t$, which is the probability generating function of the deterministic distribution ($P_x = 1,\ P_c \in R$), one obtains the Poisson distribution. And if one sets $a = -k \cdot \ln p$ and $H(t) = \ln(1-qt)/\ln(1-q)$, which is the probability generating function of the logarithmic distribution, then one obtains the negative binomial distribution applied by Grotjahn. However, both distributions can also

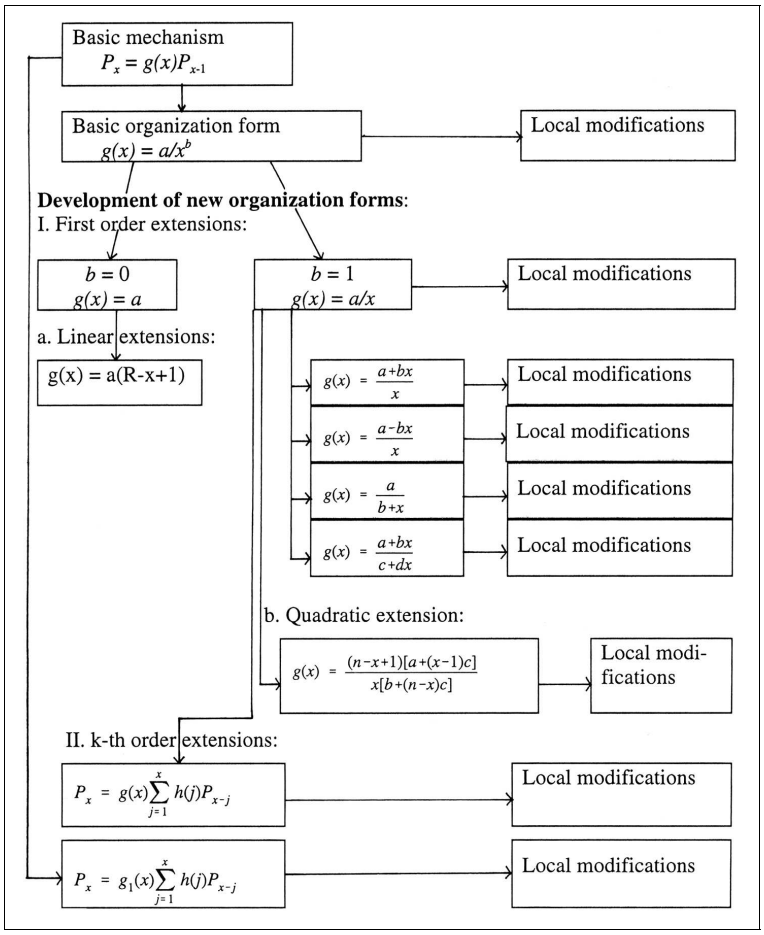


Figure 2.14: Modifications of Frequency Distributions (Wimmer/Altmann 1996)

(and more easily) be derived directly from (2.42), as was already mentioned above.

In their subsequent article on "The Theory of Word Length", Wimmer/Altmann (1996) then elaborated on their idea of different-order extensions and modifications of the postulated basic mechanism and the basic organization form resulting from it. Figure 2.14, taken from Wimmer/Altmann (1996: 114), illustrates the complete schema.

It would go beyond the frame of the present article to discuss the various extensions and modifications in detail here. In fact, Wimmer/Altmann (1996) have not only discussed the various extensions, as shown in Figure 2.14; they have also shown which concrete distributions result from these modifications.

Furthermore, they have provided empirical evidence for them from various analyses, involving different languages, authors, and texts, etc.

As a result, there seems to be increasing reason to assume that there is indeed no unique overall distribution which might cover all linguistic phenomena; rather, different distributions may be adequate with regard to the material studied. This assumption has been corroborated by a lot of empirical work on word length studies from the second half of the 1990s onwards. This work is best documented in the ongoing "Göttingen Project", managed by Best (cf. `http://wwwuser.gwdg.de/~kbest/projekt.htm`), and his bibliography (cf. Best 2001).

More often than not, the relevant analyses have been made with specialized software, usually the ALTMANN FITTER. This is an interactive computer program for fitting theoretical univariate discrete probability functions to empirical frequency distributions; fitting starts with the common point estimates and is optimized by way of iterative procedures.

There can be no doubt about the merits of such a program. Previous, deductive approaches with particular a priori assumptions dominated studies on word length, beginning with Elderton's work. Now, the door is open for inductive research, too, and the danger of arriving at ad-hoc solutions is more virulent than ever before. What is important, therefore, at present, is an abductive approach which, on the one hand, has theory-driven hypotheses at its background, but which is open for empirical findings which might make it necessary to modify the theoretical assumptions.

With this in mind, it seems worthwhile to apply this procedure once again to the Fucks' data from Table 2.9. Now, as opposed to previous approaches, we will not only go the inductive way, but we will also see how the result(s) obtained related to Wimmer/Altmann's (1994, 1996) theoretical assumptions outlined above.

Table 2.23 represents the results for that distribution which was able to model the data of all nine languages, and which, in this sense, yielded the best fitting values: we are concerned with the so-called hyper-Poisson distribution, which has two parameters (a and b). In addition to the C values of the discrepancy coefficient, the values for parameters a and b (as a result of the fitting) are given.

As can be seen, fitting results are really good in all cases. As to the data analyzed, at least, the hyper-Poisson distribution should be taken into account as an alternative model, in addition to the GDP, suggested by Stadlober (2003). Comparing these two models, a great advantage of the GPD is the fact that its reference value can be very easily calculated – this is not so convenient in the case of the hyper-Poisson distribution. On the other hand, the generation of the hyper-Poisson distribution does not involve any secondary distribution to come into play; rather, it can be directly derived from equation (2.42). Let us therefore discuss the hyper-Poisson distribution in terms of the suggestions

Table 2.23: Fitting the Hyper-Poisson Distribution to Fucks' Data From Nine Languages

	English	**German**	**Esperanto**	**Arabic**	**Greek**
$\hat{a}$	60.7124	1.1619	0.8462	0.5215	1.9095
$\hat{b}$	207.8074	2.1928	0.9115	0.2382	2.2565
C	0.0024	0.0028	0.0022	0.0068	0.0047

	Japanese	**Russian**	**Latin**	**Turkish**
$\hat{a}$	1.8581	1.8461	1.2360	1.0875
$\hat{b}$	2.1247	1.9269	0.7904	0.5403
C	0.0069	0.0029	0.0152	0.0023

made by Wimmer et al. (1994), and Wimmer/Altmann (1996), respectively. As was mentioned above, the hyper-Poisson distribution can be understood to be a "first-order extension" of the basic organization form $g(x) = a/x^b$: Setting $b = 1$, in (2.43), the corresponding extension has the form $g(x) = a/(c + x)$, which, after re-parametrization, leads to the hyper-Poisson distribution:

$$P_x = \frac{a^x}{{}_1F_1(1; b; a) \cdot b^{(x)}} , \qquad x = 0, 1, 2, \ldots \qquad a \geq 0, b > 0 . \tag{2.52}$$

Here, ${}_1F_1(1; b; a)$ is the confluent hypergeometric function

$${}_1F_1(1; b; a) = \sum_{j=0}^{\infty} \frac{a^j}{b^{(j)}} = 1 + \frac{a^1}{b^{(1)}} + \frac{a^2}{b^{(2)}} + \ldots$$

and

$$b^{(0)} = 1 ,$$
$$b^{(j)} = b\,(b + 1)\,(b + 2) \ldots (b + j - 1).$$

In its 1-displaced form, equation (2.52) takes the following shape:

$$P_x = \frac{a^{x-1}}{{}_1F_1(1; b; a) \cdot b^{(x-1)}} , \qquad x = 1, 2, 3, \ldots \qquad a \geq 0, b > 0 . \tag{2.52a}$$

As can be seen, if $b = 1$ in equation (2.52) or (2.52a), respectively, we obtain the ordinary Poisson distribution (2.8); also, what is relevant for the

English data, if $a \to \infty, b \to \infty$, and $a/b \to q$, one obtains the geometric distribution (2.1), or (2.2), respectively.

To summarize, we can thus state that the synergetic approach as developed by Wimmer et al. (1994) and Wimmer/Altmann (1996), has turned out to be extremely fruitful over the last years, and it continues to be so still today. Much empirical research has thus been provided which is in agreement with the authors' hypothesis as to a basic organization form from which, by way of extension and modification[14], further distribution models can be derived.

Most recently, Wimmer/Altmann (2005) have presented an approach which provides an overall unification of linguistic hypotheses. Generally speaking, the authors understand their contribution to be a logical extension of their synergetic approach, unifying previous assumptions and empirical findings. The individual hypotheses belonging to the proposed system have been set up earlier; they are well-known from empirical research of the last decades, and they are partly derived from different approaches.

In this approach, Wimmer/Altmann start by setting up a relative rate of change saying what should be the first step when dealing with discrete variables. According to their suggestions, this rate of change should be based on the difference $\Delta x = x - (x-1) = 1$, and consequently has the general form

$$\frac{\Delta P_{x-1}}{P_{x-1}} = \frac{P_x - P_{x-1}}{P_{x-1}}. \tag{2.53}$$

According to Wimmer/Altmann (2005), this results in the open equation

$$\frac{\Delta P_{x-1}}{P_{x-1}} = a_0 + \sum_{i=1}^{k_1} \frac{a_{1i}}{(x - b_{1i})^{c_1}} + \sum_{i=1}^{k_2} \frac{a_{2i}}{(x - b_{2i})^{c_2}} + ... \tag{2.54}$$

Now, from this general formula (2.54), different families of distributions may be derived, representing an overall model depending on the (linguistic) material to be modelled, or, mathematically speaking, depending on the definition of the parameters involved. If, for example, $k_1 = k_2 = \ldots = 1, b_{11} = b_{21} = \ldots = 0, c_i = i, a_{i1} = a_i, i = 1, 2, \ldots$, then one obtains formula (2.55), given by Wimmer/Altmann (2005):

$$P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots\right) P_{x-1}\ . \tag{2.55}$$

As to concrete linguistic analyses, particularly relevant for word length studies, the most widely used form at present seems to be (2.56). As can be seen,

[14] The authors have discussed further so-called "local" modifications, which need not be discussed here. Specifically, Wimmer et al. (1999) have discussed the modification of probability distributions, applied to Word Length Research, at some length.

it is confined to the first four terms of formula (2.54), with $k_1 = k_2 = \ldots = 1, c_i = 1, a_{i1} = a_i, b_{i1} = b_i, i = 1, 2, \ldots$. Many distributions can be derived from (2.54), which have frequently been used in linguistics studies, and which are thus united under one common roof:

$$P_x = \left(1 + a_0 + \frac{a_1}{x - b_1} + \frac{a_2}{x - b_2}\right) P_{x-1} \ . \tag{2.56}$$

Let us, in order to arrive at an end of the history and methodology of word length studies, discuss the relevant distributions discussed before, on the background of these theoretical assumptions.

Thus, for example, with $-1 < a_0 < 0$, $a_i = 0$ for $i = 1, 2, \ldots$, one obtains from (2.56)

$$P_x = (1 + a_0) P_{x-1} \tag{2.57}$$

resulting in the geometric distribution (with $1 + a_0 = q, 0 < q < 1, p = 1 - q$) in the form

$$P_x = p \cdot q^x \ , \quad x = 0, 1, 2, \ldots \tag{2.58}$$

Or, for $-1 < a_0 < 0$, $-a_1 < 1 + a_0$ and $a_2 = b_1 = b_2 = 0$, one obtains from (2.56)

$$P_{x+1} = \frac{1 + a_0 + a_1 + (1 + a_0)x}{x + 1} P_x \ . \tag{2.59}$$

With $k = (1 + a_0 + a_1)/(1 + a_0), p = -a_0$, and $q = 1 - p$ this leads to the negative binomial distribution:

$$P_x = \left(\begin{array}{c} k + x - 1 \\ x \end{array} \right) p^k q^x \ , \quad x = 0, 1, 2, \ldots \tag{2.60}$$

Finally, inserting $a_2 = 0$ in (2.56), one obtains

$$P_x = \frac{(1 + a_0)(x - b_1) + a_1}{x - b_1} P_{x-1} \tag{2.61}$$

from which the hyper-Poisson distribution (2.52) can be derived, with $a_0 = -1, b_1 = 1 - b, a_1 = a \geq 0$, and $b > 0$.

It can thus be said that the general theoretical assumptions implied in the synergetic approach has experienced strong empirical support. One may object that this is only one of possible alternative models, only one theory among others. However, thus far, we do not have any other, which is as theoretically sound, and as empirically supported, as the one presented.

It seems to be a historical absurdity, therefore, that the methodological discussion on word length studies, which was initiated by Grotjahn/Altmann (1993)

about a decade ago, has often not been sufficiently taken account of in the relevant research: more often than not, research has concentrated on word length models for particular languages, not taking notice of the fact that boundary and subsidiary conditions of individual text productions may be so strong that no overall model is adequate, not even within a given language. On the other hand, hardly any systematic studies have been undertaken to empirically study possible influencing factors, neither as to the data basis in general (i.e., text, text segments, mixtures, etc.), nor as to specific questions such as authorship, text type, etc.

Ultimately, the question, what may influence word length frequencies, may be a bottomless pit – after all, any text production is an historically unique event, the boundary conditions of which may never be reproduced, at least not completely. Still, the question remains open if particular factors may be detected, the relevance of which for the distribution of word length frequencies may be proven.

This point definitely goes beyond a historical survey of word length studies; rather, it directs our attention to research desires, as a result of the methodological discussion above. As can be seen, the situation has remained unchanged: in this respect, it will always be a matter of orientation, or of object definition, if one attempts to find "local" solutions (on the basis of a clearly defined data basis), or general solutions, attempting a general explanation of language or text processing.

References

Altmann, Gabriel; Hammer, Rolf
1989 *Diskrete Wahrscheinlichkeitsverteilungen I.* Bochum.

Antić, Gordana; Grzybek, Peter; Stadlober, Ernst
2005a "Mathematical Aspects and Modifications of Fucks' Generalized Poisson Distribution." In: Köhler, R.; Altmann, G.; Piotrovskij, R.G. (eds.), *Handbook of Quantitative Linguistics.* [In print]

Antić, Gordana; Grzybek, Peter; Stadlober, Ernst
2005b "50 Years of Fucks' 'Theory of Word Formation': The Fucks Generalized Poisson Distribution in Theory and Praxis." In: *Journal of Quantitative Linguistics.* [In print]

Bagnold, R.A.
1983 "The nature and correlation of random distributions", in: *Proceedings of the Royal Society of London, ser. A, 388*; 273–291.

Bartkowiakowa, Anna; Gleichgewicht, Bolesław
1962 "O długości sylabicznej wyrazów w tekstach autorów polskich", in: *Zastosowania matematyki, 6*; 309–319. [= On the syllable length of words in texts by Polish authors]

Bartkowiakowa, Anna; Gleichgewicht, Bolesław
1964 "Zastosowanie dwuparametrowych rozkładów Fucksa do opisu długości sylabicznej wyrazów w różnych utworach prozaicznych autorów polskich", in: *Zastosowania matematyki, 7*; 345–352. [= Application of two-parameter Fucks distributions to the description of syllable length of words in various prose texts by Polish authors]

Bartkowiakowa, Anna; Gleichgewicht, Bolesław
1965 "O rozkładach długości sylabicznej wyrazów w różnych tekstach." In: Mayenowa, M.R. (ed.), *Poetyka i matematyka.* Warszawa. (164–173). [= On the distribution of syllable length of words in various texts.]

Best, Karl-Heinz
1997 "Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten." In: Best, K.-H. (ed.), *Glottometrika 16: The Distribution of Word and Sentence Length.* Trier. (1–15).

Best, Karl-Heinz; Čebanov, Sergej G.
2001 "Biographische Notiz: Sergej Grigor'evič Čebanov (1897–1966)." In: Best (ed.) (2001); 281–283.

Best, Karl-Heinz
2001 "Kommentierte Bibliographie zum Göttinger Projekt." In: Best, K.-H. (ed.) (2001); 284–310.

Best, Karl-Heinz (ed.)
2001 *Häufigkeitsverteilungen in Texten.* Göttingen.

Brainerd, Barron
1971 "On the distribution of syllables per word", in: *Mathematical Linguistics [Keiryo Kokugogaku], 57*; 1–18.

Brainerd, Barron
1975 *Weighing evidence in language and literature: A statistical approach.* Toronto.

Cercvadze, G.N.; Čikoidze, G.B./Gačečiladze, T.G.
1959 "Primenenie matematičeskoj teorii slovoobrazovanija k gruzinskomu jazyku", in: *Soobščenija akademii nauk Gruzinskoj SSR, t. 22/6*, 705–710.

Cercvadze, G.N.; Čikoidze, G.B./Gačečiladze, T.G.
1962 see: Zerzwadse et al. (1962)

Čebanov s. Chebanow

Chebanow, S.G.
1947 "On Conformity of Language Structures within the Indo-European Familiy to Poisson's Law", in: *Comptes Rendus (Doklady) de l'Adadémie des Sciences de l'URS, vol. 55, no. 2*; 99–102.

Dewey, G.
1923 *Relative Frequencies of English Speech Sounds.* Cambridge; Mass.

Elderton, William P.
1949 "A Few Statistics on the Length of English Words", in: *Journal of the Royal Statistical Society, series A (general), vol. 112*; 436–445.

French, N.R.; Carter, C.W.; Koenig, W.
1930 "Words and Sounds of Telephone Communications", in: *Bell System Technical Journal, 9*; 290–325.

Fucks, Wilhelm
1955a *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln/Opladen. [= Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen; 34a]

Fucks, Wilhelm
1955b "Theorie der Wortbildung", in: *Mathematisch-Physikalische Semesterberichte zur Pflege des Zusammenhangs von Schule und Universität, 4*; 195–212.

Fucks, Wilhelm
1955c "Eine statistische Verteilung mit Vorbelegung. Anwendung auf mathematische Sprachanalyse", in: *Die Naturwissenschaften, 42*$_1$; 10.

Fucks, Wilhelm
1956a "Die mathematischen Gesetze der Bildung von Sprachelementen aus ihren Bestandteilen", in: *Nachrichtentechnische Fachberichte, 3 [= Beiheft zu Nachrichtentechnische Fachzeitschrift]*; 7–21.

Fucks, Wilhelm
1956b "Mathematische Analyse von Werken der Sprache und der Musik", in: *Physikalische Blätter, 16*; 452–459 & 545.

Fucks, Wilhelm
1956c "Statistische Verteilungen mit gebundenen Anteilen', in: *Zeitschrift für Physik, 145*; 520–533.

Fucks, Wilhelm
1956d "Mathematical theory of word formation." In: Cherry, Colin (ed.), *Information theory.* London, 1955. (154–170).

Fucks, Wilhelm
1957 "Gibt es allgemeine Gesetze in Sprache und Musik?", in: *Umschau, 57*$_2$; 33–37.

Fucks, Wilhelm
1960 "Mathematische Analyse von Werken der Sprache und der Musik", in: *Physikalische Blätter, 16*; 452–459.

Fucks, Wilhelm; Lauter, Josef
1968 "Mathematische Analyse des literarischen Stils." In: Kreuzer, H.; Gunzenhäuser, R. (eds.), *Mathematik und Dichtung.* München, ^{4}1971.

Fucks, Wilhelm
1968 *Nach allen Regeln der Kunst. Diagnosen über Literatur, Musik, bildende Kunst – die Werke, ihre Autoren und Schöpfer.* Stuttgart.

Gačečiladze, T.G.; Cilosani, T.P.
1971 "Ob odnom metode izučenija statističeskoj struktury teksta." In: *Statistika reči i avtomatičeskij analiz teksta*. Leningrad, Nauka: 113–133.

Grotjahn, Rüdiger
1982 "Ein statistisches Modell für die Verteilung der Wortlänge", in: *Zeitschrift für Sprachwissenschaft, 1*; 44–75.

Grotjahn, Rüdiger; Altmann, Gabriel
1993 "Modelling the Distribution of Word Length: Some Methodological Problems." In: Köhler, R.; Rieger, B. (eds.), *Contributions to Quantitative Linguistics.* Dordrecht, NL. (141–153).

Grzybek, Peter
2001 "Pogostnostna analiza besed iz elektronskego korpusa slovenskih besedel", in: *Slavistična Revija, 48(2) 2000 [2001]*; 141–157.

Grzybek, Peter (ed.)
2004 *Studies on the Generalized Fucks Model of Word Length.* [In prep.]

Grzybek, Peter; Kelih, Emmerich; Altmann, Gabriel
2004 "Graphemhäufigkeiten (Am Beispiel des Russischen) Teil II: Theoretische Modelle", in: *Anzeiger für Slavische Philologie, 32*; 25–54.

Grzybek, Peter; Kelih, Emmerich
2005 "Texttypologie in/aus empirischer Perspektive." In: Bernard, J.; Fikfak, J.; Grzybek, P. (eds.), *Text und Realität – Text and Reality.* Ljubljana etc. (95–120).

Grzybek, Peter; Stadlober, Ernst
2003 "Zur Prosa Karel Čapeks – Einige quantitative Bemerkungen." In: Kempgen, S.; Schweier, U.; Berger, T. (eds.), *Rusistika • Slavistika • Lingvistika.* München. (474–488).

Grzybek, Peter; Stadlober, Ernst; Antić, Gordana; Kelih, Emmerich
2005 "Quantitative Text Typology: The Impact of Word Length." In: Weihs, C. (ed.), *Classification – The Ubiquitous Challenge.* Heidelberg/Berlin. (53–64).

Herdan, Gustav
1958 "The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics", in: *Biometrika, 45*; 222–228.

Herdan, Gustav
1966 *The Advanced Theory of Language as Choice and Chance.* Berlin etc.

Kelih, Emmerich; Antić, Gordana; Grzybek, Peter; Stadlober, Ernst
2005 "Classification of Author and/or Genre? The Impact of Word Length." In: Weihs, C. (ed.), *Classification – The Ubiquitous Challenge.* Heidelberg/Berlin. [In print]

Kromer, Victor V.
2001a "Word length model based on the one-displaced Poisson-uniform distribution", in: *Glottometrics, 1*; 87–96.

Kromer, Victor V.
2001b "Dvuchparametričeskaja model' dliny slova 'jazyk –žanr' . [= A Two-Parameter Model of Word Length: ≪Language – Genre≫.]" In: *Electronic archive* Computer Science, March 8, 2001 [http://arxiv.org/abs/cs.CL/0103007].

Kromer, Victor V.
2001c "Matematičeskaja model' dliny slova na osnove raspredelenija Čebanova-Fuksa s ravnomernym raspredeleniem parametra. [= A Mathematical Model of Word Length on the Basis of the Čebanov-Fucks Distribution with Uniform Distribution of the Parameter.]" In: *Informatika i problemy telekommunikacij: meždunarodnaja naučno-tekhničeskaja konferencija SibGUTI, 26-27 aprelja 2001 g. Materialy konferencii.* Novosibirsk. (74–75). [http://kromer.newmail.ru/kvv_c_18.pdf]

Kromer, Victor V.
2002 "Ob odnoj vozmožnosti obobščenija matematičeskoj modeli dliny slova. [= On A Possible Generalization of the Word Length Model.]" In: *Informatika i problemy telekommunikacij: meždunarodnaja naučno-tekhničeskaja konferencija SibGUTI, 25-26 aprelja 2002 g. Materialy konferencii.* Novosibirsk. (139–140). [http://kromer.newmail.ru/kvv_c_23.pdf]

Lord, R.D.
1958 "Studies in the history of probability and statistics. VIII: De Morgan and the statistical study of literary style", in: *Biometrika, 45*; 282.

Markov, Andrej A.
1924 *Isčislenie verojatnostej.* Moskva.

Mendenhall, Thomas C.
1887 "The characteristic curves of composition", in: *Science, supplement, vol. 214, pt. 9*; 237–249.

Mendenhall, Thomas C.
1901 "A mechanical solution of a literary problem", in: *Popular Science Monthly, vol. 60, pt. 7*; 97–105.

Merkytė, R.Ju.
1972 "Zakon, opisyvajuščij raspredelenie slogov v slovach slovarej", in: *Lietuvos matematikos rinkinys, 12/4*; 125–131.

Michel, Gunther
1982 "Zur Häufigkeitsverteilung der Wortlänge im Bulgarischen und im Griechischen." In: *1300 Jahre Bulgarien. Studien zum 1. Internationalen Bulgaristikkongress Sofia 1981.* Neuried. (143–208).

Moreau, René
1961 "Linguistique quantitative. Sur la distribution des unités lexicales dans le français écrit", in: *Comptes rendus hebdomaires des séances de l'académie des sciences, 253*; 2626–2628.

Moreau, René
1963 "Sur la distribution des formes verbales dans le français écrit", in: *Études de linguistique appliquée, 2*; 65–88.
Piotrovskij, Rajmond G.; Bektaev, Kaldybaj B.; Piotrovskaja, Anna A.
1977 *Matematičeskaja lingvistika.* Moskva. [German translation: Piotrowski, R.G.; Bektaev, K.B.; Piotrowskaja, A.A.: *Mathematische Linguistik.* Bochum, 1985.]
Rothschild, Lord
1986 "The Distribution of English Dictionary Word Lengths", in: *Journal of Statistical Planning and Inference, 14*; 311–322.
Stadlober, Ernst
2003 "Poissonmodelle und Wortlängenhäufigkeiten." [Ms.]
Vranić, V.
1965a "Statističko istraživanje hrvatskosrpskog jezika", in: *Statistička revija, 15(2-3)*; 174–185.
Vranić, V.; Matković, V.
1965b "Mathematic Theory of the Syllabic Structure of Croato-Serbian", in: *Rad JAZU (odjel za matematičke, fizičke i tehničke nauke; 10) (331)*; 181–199.
Williams, Carrington B.
1939 "A note on the statistical analysis of sentence-length as a criterion of literary style", in: *Biometrika, 31*; 356–361.
Williams, Carrington B.
1956 "Studies in the history of probability and statistics. IV: A note on an early statistical study of literary style", in: *Biometrika, 43*; 248–256.
Williams, Carrington B.
1967 "Writers, readers and arithmetic", in: *New Scientist, 13*; 88–91.
Williams, Carrington B.
1976 "Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon", in: *Biometrika, 62*; 207–212.
Wimmer, Gejza; Altmann, Gabriel
1996 "The Theory of Word Length: Some Results and Generalizations." In: Schmidt, P. (ed.), *Glottometrika 15: Issues in General Linguistic Theory and the Theory of Word Length.* Trier. (112–133).
Wimmer, Gejza; Altmann, Gabriel
1999 *Thesaurus of univariate discrete probability distributions.* Essen.
Wimmer, Gejza; Altmann, Gabriel
2005 "Unified derivation of some linguistic laws." In: Köhler, R.; Altmann, G.; Piotrovskij, R.G. (eds.), *Handbook of Quantitative Linguistics.* [In print]
Wimmer, Gejza; Köhler, Reinhard; Grotjahn, Rüdiger; Altmann, Gabriel
1994 "Towards a Theory of Word Length Distribution", in: *Journal of Quantitative Linguistics, 1/1*; 98–106.
Wimmer, Gejza; Witkovský, Viktor; Altmann, Gabriel
1999 "Modification of Probability Distributions Applied to Word Length Research", in: *Journal of Quantitative Linguistics, 6/3*; 257–268.
Zerzwadse, G.; Tschikoidse, G.; Gatschetschiladse, Th.
1962 Die Anwendung der mathematischen Theorie der Wortbildung auf die georgische Sprache. In: *Grundlagenstudien aus Kybernetik und Geisteswissenschaft* 4, 110–118.
Ziegler, Arne
1996 "Word Length Distribution in Brazilian-Portuguese Texts", in: *Journal of Quantitative Linguistics,* 3_1; 73–79.
Ziegler, Arne
1996 "Word Length in Portuguese Texts", in: *Journal of Quantitative Linguistics,* 5_{1-2}; 115–120.