REGULARITIES OF ESTONIAN PROVERB WORD LENGTH: FREQUENCIES, SEQUENCES, DEPENDENCIES

Peter Grzybek

University of Graz, Austria

Abstract

Concentrating on specifics of word length in Estonian proverbs, this article is part of a more comprehensive analysis of regularities characterizing their linguistic organization. Based on empirical data provided and published by Arvo Krikmann in his 1967 study *Keelestatistikat Eesti vanasõnadest* (Linguostatistics of Estonian proverbs), an attempt is made to arrive at generalizing conclusions by discussing theoretical models for tendencies emerging from Krikmann's data. In detail, models of word length frequencies, as well as models for the relationship between word length and sentence length, and for the relationship between word length and position within a proverbial sentence are discussed.

Keywords: frequencies, proverb, sentence length, sequences, word length

Introduction

Over the years, it has become a truism to say that the proverb is the property of the folk. And it is also commonplace to state that, although we do have the discipline of paremiology, the proverb may be (and has been) studied from many kinds of scholarly discipline, starting from folkloristics, through sociology, to pedagogics, and many others, each of them having a specific perspective and asking different questions. As a matter of fact, with the proverb being part of verbal folklore, the discipline of linguistics has also been traditionally concerned; however, as in linguistics in general, linguistic studies of the proverb tend to be characterized by symptomatic, rather than systematic, approaches, focusing either on isolated phenomena or specifically selected ('demonstrative') material.

Systematic approaches, searching for fundamental regularities of the linguistic organization of proverbs, are quite uncommon and represent a scientific deficit. In this respect, Arvo Krikmann's study *Keelestatistikat Eesti vanasõnadest* (Linguostatistics of Estonian Proverbs; 1967) is a remarkable exception: it is one of the earliest and, in fact, one of the few quantitative studies of proverbs in general. Unfortunately, this study has, except for a short abstract in Russian and a short summarizing communication in German by Pentti Leino (1968), never been translated from Estonian[1]. This seems to be the reason why it has remained almost unknown to the international academic world, although, at the time of its publication, it was much ahead of its time. Because of its systematic approach, it is still today appropriate to serve as a source of inspiration and as a starting point, and also for comparative studies that attempt to achieve regularities in the linguistic organization of proverbs.

The present contribution intends to recall the cornerstones of those achievements, made some decades ago. Moreover, attempts shall be made to re-analyze some major findings from a contemporary point of view, based on insights from the field of quantitative linguistics achieved in the last decades, thus paving the way to draw generalizations from the results obtained.

Krikmann based his analyses primarily on Erna Normann's collection *Valimik eesti vanasõnu* (A Selection of Estonian Proverbs; 1955), which contains 3,576 items from the end of the 19th and early 20th centuries. On the basis of this proverbial material, he concentrated on sentence and word length in Estonian proverbs by asking the following four questions:

a. Is there a regularity of sentence length frequencies in proverbs, i.e., do proverbs of a given length occur with arbitrary (or random) frequency, or is there a specific regularity to these frequencies?
b. Is there a regularity of word length frequencies in the proverbs, i.e., do words of a given length occur arbitrarily, or is there a specific regularity to these frequencies?
c. Is there a specific relationship between sentence length and word length in proverbs?
d. Are there positional regularities of word length in proverbs, i.e., are there specific regularities in the sequence of words of given length that may go along with specific rhythmic regularities?

[1] The author of this contribution does not want to give rise to the false impression that he sufficiently understands Kriq's original Estonian text. Rather, Kriq was so kind as to translate the gist of his study into English in a letter from November 26, 1999. At that time, there had been contacts between us two already for more than 15 years, although we were to meet personally for the first time only in August 2008.

As has been pointed out at the beginning, the present contribution will concentrate on word length, that is, on the last three points mentioned above.

Word length

As is well known from the field of quantitative linguistics today, word length is not an isolated category within a linguistic system, but closely interrelated to other properties of the word, as well as of other linguistic units, levels, and structures. In addition, are the frequencies and manners in which words of a given length occur in linguistic material not chaotic, but rather do they follow clearly defined, law-like regularities? Most of these regularities were not yet known in the 1960s, when most approaches were of merely empirical orientation and emerged as ad-hoc solutions to 'local' problems rather than against the background of a comprehensive theory.

When measuring the length of linguistic phenomena it is necessary to define both the unit to be measured and the measuring unit. Krikmann's approach corresponds to standards in quantitative linguistics that are still common today, though not identical across languages: a word is defined as an orthographic unit, and word length is measured by the number of syllables per word.[2]

Under these conditions, the word length data in Table 1 are obtained for Normann's proverbs, x denoting a given word length class, f_x the corresponding frequencies of occurrence.

Table 1. *Word length frequencies.*

x	f_x
1	6,648
2	10,573
3	2,730
4	920
5	149
6	16
7	2

[2] In the tradition of Walter Anderson (1935), compounds are counted as one word in Krikmann's study.

The frequencies are graphically illustrated in form of a bar chart in Figure 1.

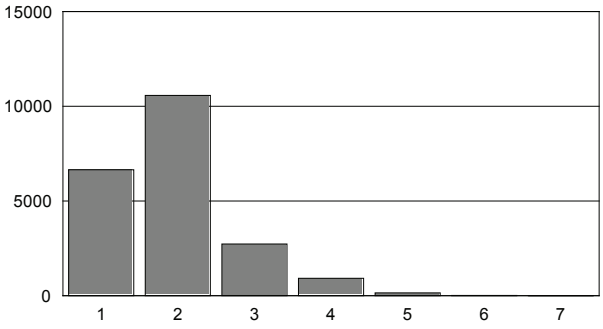


Figure 1. *Word length frequencies.*

From these data, descriptive statistics can be derived to characterize the frequency distribution. Such descriptive measures would be, among others, mean value, variance or standard deviation, skewness, kurtosis, entropy, repeat rate, or many others. In our context, we do not need most of them; let it therefore suffice to say that on the basis of our sample size of $N = 21{,}038$ words, a mean word length of $\bar{x} = 1.9260$ syllables per word is obtained, with a standard deviation of $s = 0.83$.

The question of whether there is a regularity in word length frequency distribution goes beyond descriptive statistics, since in this case we are concerned with looking for a model by which the frequencies can be described, and which may serve as the basis for comparisons with other frequency distributions. In finding such a model, one may in principle choose between a continuous and a discrete approach, the first usually represented by continuous functions, the second by discrete probability distributions. Since in our case, we are concerned with discrete units of $x = 1, 2, 3$, etc. syllables per word, it is reasonable to search for a discrete model.

Such models do not, of course, fall from the sky, and they are not God-given truths. Rather, they result from some generating process. In contemporary quantitative linguistics, it has become common to understand the frequencies of a given distribution to be mutually dependent; in detail, the frequency of a given class x_i is seen in relation to its preceding class x_{i-1}. In other words, the frequency of three-syllable words, for example, is not independent of the number of two-syllable words, and the number of two-syllable words is related to the number of one-syllable words. Mathematically, this is expressed in terms of

a function $g(x)$, where the probability P_x of a given class is related to that of the preceding class P_{x-1}:

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

Depending on the form of function $g(x)$, different frequency models and, as a result, different frequencies, are obtained. Setting, for example, $g(x) = a/x$ we obtain the difference equation

$$P_x = \frac{a}{x} P_{x-1} \tag{2}$$

which results in the well-known Poisson distribution:

$$P_x = \frac{e^{-a} a^x}{x!} \qquad x = 0,1,2, \ldots \tag{3}$$

As can be seen, the Poisson distribution has one parameter (a), and depending on the parameter value of a, the frequencies may differ, although we are still concerned with one and the same model. This parameter value thus needs to be estimated, to fit the model to the observed data, and then to test the how good the fit is by way of a statistical test. As a result of this test, the model can either be retained if the differences are statistically random, or it must be rejected if the differences are statistically significant.

Thus, in his approach, Krikmann was not satisfied by providing empirical data and descriptive statistics, as his aim was to find a theoretical (probabilistic) model for the observed frequencies. Moreover, he was thus fully in line with quantitative research of his time when he attempted to fit the 1-parameter Poisson distribution to the observed data; since there are no 0-syllable words in Estonian (i.e., minimal word length is $x = 1$, so that $f_0 = 0$), he used the Poisson distribution in its 1-shifted form, which at that time was widely known as 'Fucks distribution'.[3]

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

[3] In fact, the Fucks distribution is a more complex model, representing a generalization of the Poisson distribution with specific weights (cf. Antić et al. 2005); details can be ignored in this context, however.

There are many different methods for parameter estimation: whereas in former times, estimation methods were applied which were derived from theoretical considerations, today specialized computer programs are (additionally) available which contain specific algorithms for iterative procedures. In our case, iterative procedures provide the same parameter value for a as estimating it by the distribution's mean, as is possible in the case of the Poisson distribution – i.e., $a = \bar{x} = 1.9260$. As a result, the theoretical values represented in Table 2 are obtained.

Table 2. *Empirical and theoretical word length frequencies (Poisson distribution).*

x	f_x	Np_x
1	6,648	8,333.96
2	10,573	7,717.18
3	2,730	3,573.02
4	920	1,102.86
5	149	255.31
6	16	47.28
7	2	7.30

Figure 2 offers a graphical comparison of both observed (f_x) and theoretical (Np_x) values.

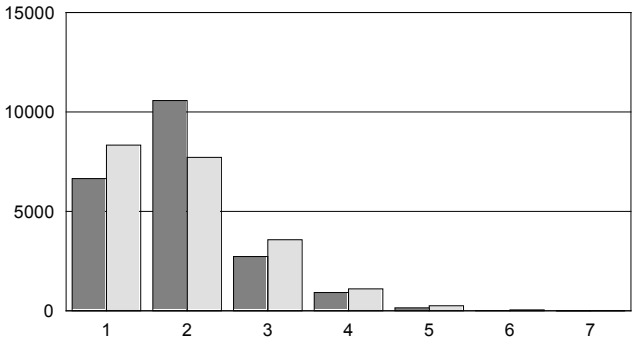


Figure 2. *Empirical and theoretical word length frequencies (Poisson distribution).*

As can be seen immediately, the theoretical frequencies do not appear to correspond to the observed ones to a satisfying degree. In order to objectivize this impression, statistical goodness-of-fit tests are usually run today, which was less common at that time. Today, it would be usual to apply the well-known chi square test for this purpose. From a practical perspective, this test has, however, a major disadvantage: the chi square value linearly increases with an increase in sample size; however, in the case under study here, as in linguistics in general, we are used to have rather large samples, so differences between observed and theoretical frequencies tend to become significant sooner. In order to minimize this problem, it has become common to use the standardized discrepancy coefficient $C = X^2/N$, with a value of $C < 0.02$ being interpreted as a good, of $C < 0.01$ as a very good fit.

In the case of the (1-shifted) Poisson distribution we obtain a value of $C = 0.08$ for the data presented here, which is indeed far from satisfactory. As a consequence, the Poisson model must be rejected as an inadequate model for the word length frequency distribution in the chosen proverbs.

However, the question of modelling word length frequencies has been a major topic in quantitative linguistics in recent years (cf. Grotjahn & Altmann 1993; Wimmer et al. 1994; Grzybek 2006; Grzybek 2014). In this context it has been shown that there is not, as has previously been assumed, one common ('universal') model for word length frequencies from different languages; rather, there is a general generating mechanism (cf. Wimmer & Altmann 2005; 2006) from which various models can be derived, depending on the language analyzed, and allowing for further modifications or specifications, due to particular boundary conditions (i.e., factors such as language, text type, and others).

This discussion needs not be presented here in detail. Suffice it to say that for the frequency distribution of word length in Estonian, a generalization of the Poisson distribution has been suggested (Bartens & Best) to be a good model, specifically the 2-parameter (a, b) hyper-Poisson distribution (cf. Wimmer & Altmann 1999: 281f.) in its 1-shifted form (4):[4]

$$P_x = \frac{a^{x-1}}{{}_1F_1(1;b;a)\,b^{(x-1)}} \qquad x = 0, 1, 2, \ldots \tag{4}$$

[4] In (4), ${}_1F_1$ is the confluent hypergeometric function, and $b^{(x-1)}$ denotes the ascending factorial function $b(b+1)(b+2)\ldots(b+x-2)$.

Applying this model[5] to the data collected, it turns out that it indeed yields more appropriate results, with a satisfying discrepancy coefficient value of $C = 0.0152$ ($a = 0.48$, $b = 0.33$). We may assume, therefore, that the organization of word length in our proverbs is not random or arbitrary, but follows specific regularities, which seems to be in line with findings for other Estonian texts.

In an earlier re-analysis of the Estonian proverb data (cf. Grzybek 2000), another model was suggested, providing even better results. In this case we are concerned with the 3-parameter hyper-Pascal (k, m, q) distribution (cf. Wimmer & Altmann 1999: 279ff.), again in its 1-shifted form:

$$(5) \qquad P_x = \frac{\binom{k+x-2}{x-1}}{\binom{m+x-2}{x-1}} q^{x-1} P_1 \qquad x = 1, 2, 3, \ldots$$

Having one parameter more than the hyper-Poisson distribution, the hyper-Pascal model indeed provides a much better result, with $C = 0.0062$ for parameter values $k = 0.0872$, $m = 0.0159$, and $q = 0.2674$.

As further analyses show, however, one should take into account yet another 2-parameter model, which yields similarly good results in case of our proverbs; we are concerned here with a generalization of the geometric distribution

$$(6a) \qquad P_x = pq^x \qquad \begin{array}{l} x = 0, 1, 2, \ldots \\ q = 1\text{–}p \end{array}$$

in its 1-shifted form

$$(6b) \qquad P_x = pq^{x-1} \qquad \begin{array}{l} x = 1, 2, 3, \ldots \\ q = 1\text{–}p \end{array}$$

namely, the Shenton-Skees geometric distribution (cf. Wimmer & Altmann 1999: 593), which is given as

$$(7) \qquad P_x = pq^{x-1}\left[1 + a\left(x - \frac{1}{p}\right)\right] \qquad \begin{array}{l} x = 1, 2, 3, \ldots \\ q = 1\text{–}p \end{array}$$

[5] In their analysis of 50 Estonian texts, Bartens & Best (1996) have also tested a second model, the negative binomial distribution, which, according to their findings, was less appropriate.

With parameter values $p = 0.85$ and $a = 3.59$, this model yields an excellent value of $C = 0.0062$. Table 3 represents the corresponding results.

Table 3. *Empirical and theoretical word length frequencies (Shenton-Skees geometric distribution).*

x	f_x	Np_x
1	6,648	6,519.03
2	10,573	10,628.39
3	2,730	3,048.85
4	920	676.57
5	149	134.53
6	16	25.15
7	2	5.48

Both empirical and theoretical frequencies are illustrated in Figure 3.

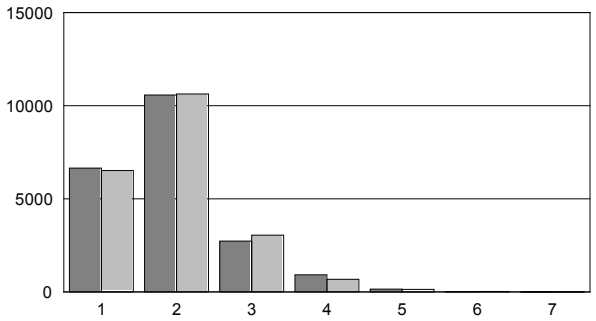


Figure 3. *Empirical and theoretical word length frequencies (Shenton-Skees geometric distribution).*

It goes without saying that when searching for an adequate model one would, or should, never favour one that (only) empirically fits, but favour one that also can be derived from theoretical ruminations: a

superior 'local' solution covering only individual data should not have an advantage over a general and theoretically justified model.[6]

In this context, one should not forget that the linguistic organization of proverbs differs from that of ordinary running texts: for example, in proverbs, there is no supra-sentential level as in coherent discourse, and all kinds of connector are absent. It may therefore well be possible that the different organization of word length in proverbs asks for a different model as compared to "ordinary" texts (in the strict sense of this word's meaning); however, in this case, the differences should also be explained in terms of the model's changes.[7]

In any case, in order to arrive at reliable results as a basis for further conclusions, the present findings suggest that word length frequencies in Estonian should systematically be re-analyzed; in this context, not only the hyper-Poisson, but also the Shenton-Skees geometric distribution should be tested for appropriateness, not only on a broader proverbial basis, but also on other kinds of running texts.[8]

With this in mind, we can next turn to the second issue raised in Krikmann's study, the relationship between a proverb's length and word length.

---

[6] In this respect, it should be mentioned that among the Estonian tests carried out by Bartens & Best (see above) there were not only poetic texts, usually characterized by a specific lexical organization, but also extremely short texts (for example, with not more than 84 words composed of only 1, 2 or 3 syllables). Moreover, even for the relatively longer texts, in most cases some classes had to be pooled to arrive at results for the hyper-Poisson distribution that fitted well. The longest text, with 1,609 words, a short story called "Elsa Hermann" from the Estonian writer Mari Saat (from *Õun valguses ja varjus* (The apple in light and shadow; 1985), consisted of 8 length classes, although here too the last four had to be pooled. Additionally, a re-analysis shows that in this case the Shenton-Skees geometric distribution turns out to be an excellent model ($C < 0.0048$) without any pooling procedures. These results clearly ask for a more comprehensive and systematic analysis of word length in Estonian, as suggested by the authors themselves, and as corroborated by my findings.

[7] Theoretically, the hyper-Poisson model converges with the geometric distribution for $a \to \infty, b \to \infty, a/b \to q$; furthermore, as can be seen above, the Shenton-Skees geometric distribution is but a 1-modified geometric distribution; however, without further analyses on broader data, any reflection in this direction is but speculation.

[8] It should not go unmentioned that for our proverbs, the Shenton-Skees distribution is the more appropriate model for word length frequencies under two specific test conditions: first, if the word length frequency distribution is calculated separately for each sentence (i.e., proverb) length, and second, for individual word positions within the proverbial texts (i.e., only for words in the first, second, third, etc. positions) – for both cases, the Shenton-Skees geometric model outranges both the hyper-Poisson and the hyper-Pascal distribution.

PROVERB LENGTH ↔ WORD LENGTH

Given that word length frequencies are regularly organized, and that there is a theoretical model to describe word length frequencies, the question of how word length is organized in the sequential order of proverbs quite naturally arises. In this respect, Krikmann's observation that again there is some regularity is of crucial importance; specifically, Krikmann observed that an increase of proverb length corresponds with a decrease in word length (1967: 138). Table 4 offers the corresponding data.

Table 4. *Mean word length for separate proverb (sentence) lengths.*

	Proverb Length							
	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
$\bar{x}$ **(Word length)**	2.2652	1.9939	1.9830	1.9554	1.9642	1.8434	1.8507	1.8217

Today we know that such "vertical" or "hierarchical" relationships between units from different (neighbouring) linguistic levels are indeed systematically organized in language. In this context, when searching for a model of this tendency, the well-known Menzerath-Altmann law is of primary relevance: generalizing previous findings by Paul Menzerath (1928) on the relationship between word length and syllable length, the formulation of this law goes back to German scholar Gabriel Altmann (1980), who generally claimed a constituent's length would decrease with an increase in the corresponding construct (for example, the longer a word, the shorter the syllables constituting this word). It should be emphasized that this tendency concerns direct relations (in the classical structuralist paradigm) only, i.e., the relationship of a construct to its immediate constituents; the relationship between entities from indirectly related levels (for example, between sentences and words, which leapfrogs the intermediate level of sub-sentential constructs like clauses or phrases) is expected to show different tendencies.

In the framework of the Menzerath-Altmann law, such relationships have frequently been modelled with the simple 2-parameter potency function

(8) $$y = K \cdot x^a,$$

where y represents the construct as the dependent variable, x the constituent as the independent variable, K the integration constant, and b a parameter determining the steepness of the decrease (for $a < 0$). This function implies the assumption that the relative change rate is characterized by an inverse proportionality, corresponding to the underlying differential equation

(8a) $$\frac{y'}{y} = \frac{a}{x}.$$

Function (8) is but a special case of the more complex Menzerathian function

(9) $$y = K \cdot x^a \cdot e^{bx},$$

based on the assumption that the simple proportionality function (8a) is not sufficient for more complex cases that ask for an additional constant to express additional monotonous decreases coming into play (for $a < 0$):

(9a) $$\frac{y'}{y} = b + \frac{a}{x}.$$

As can be seen, (8) and (8a) can be derived from the above functions (9) or (9a) respectively, for $b = 0$. More recently, both functions[9] have analogically been derived from the even more complex function

(10) $$y = K \cdot x^a \cdot e^{bx} \cdot e^{-c/x},$$

which goes back to the unified derivation of linguistic laws developed by Wimmer and Altmann (2005, 2006). Quite obviously, from (10) and its corresponding relative change rate (10a)

(10a) $$\frac{y'}{y} = b + \frac{a}{x} + \frac{c}{x^2},$$

[9] In the original (1980) version of the Menzerath-Altmann law a third function, which is not relevant in our context, was included, namely, the exponential function $y = K \cdot e^{bx}$, which is obtained from (9) for $a = 0$.

on the differential equations (8a) and (9a), can be derived. In analyses one would of course always tend to choose that function which has fewer parameters to be estimated (usually by way of iterative processes) and, more importantly, to be interpreted. With this in mind, it turns out as a result that in the case of the proverbs examined, the two-parameter function

$$y = K \cdot e^{-c/x}, \tag{11}$$

which is obtained from (8) for $a = 0$ and $b = 0$, yields a good fit: with parameter values $K = 1.68$ and $c = -0.84$, the determination coefficient is $R^2 = 0.90$. Figure 4 shows the empirical data points and the regression curve obtained from the function described.

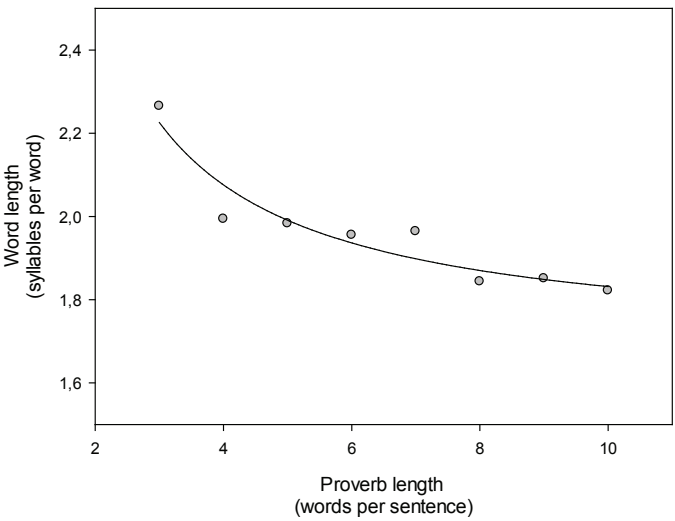


Figure 4. *Dependence of word length on proverb (sentence) length.*

This finding corroborates the assumption that there is a regular relationship between word length and proverb length. What is striking, however, is the tendency of this relationship, i.e., a decrease in word length with an increase in proverb length. Given the assumption that there is a syntactically relevant intermediate level, i.e. sub-sentential construct-like clauses, phrases, or partial sentences, a quantitative

linguistic approach would tend to regard these as immediate constituents of the sentence, and would then predict shorter sub-sentential units for longer sentences. A length decrease of the sub-sentential unit, however, should go along with an increase in word length. If both assumptions are taken together, one would expect an increase in word length with an increase in sentence (proverb) length. In fact, for running texts, such tendencies have recently been demonstrated, although not for Estonian, and with modifications for very short sentences without hypotaxis.

Again, we do not have sufficient data, at least not for Estonian, to interpret the results obtained: on the one hand, we may be concerned with proverbial specifics, eventually syntactic characteristics of the proverb; on the other hand, we may be concerned with the specifics of Estonian syntax in general. Unfortunately, we do not have any Estonian data about the relationship between word length and either sub-sentential or sentential length, and the question as to an interpretation of our findings must remain open until such data are available.

Positional aspects of word length

The next issue raised in Krikmann's study, and the last to be dealt with here, concerns the question of how far word length is characterized by particular regularities with regard to the position within a running proverbial text. Since in Normann's material we have a maximal length of x_{max}=10, ten positions (Pos_1…Pos_{10}) can be distinguished, and average word length can be calculated for each position. Based on the raw data given in Krikmann's text, the corresponding averages for each position can be calculated (cf. Table 5):

Table 5. *Mean word length for individual within-proverb positions.*

	Within-Proverb Position									
	Pos_1	**Pos_2**	**Pos_3**	**Pos_4**	**Pos_5**	**Pos_6**	**Pos_7**	**Pos_8**	**Pos_9**	**Pos_{10}**
$\bar{x}$ (Word length)	1.8852	1.7980	1.9765	1.9608	1.8943	1.9756	2.0373	1.9704	1.9771	2.1714

As can be seen on closer inspection, the mean values display some wave-like form, with minima at positions 2, 5, 8, and maxima at positions 3, 7, 10. Interestingly enough, such a tendency can adequately

be modelled in terms of an ordinary Fourier series, which decomposes periodic phenomena into the sum of a (possibly infinite) set of simple oscillating functions, namely sines and cosines. In our case, the series

$$f(x) = k + a \cdot \sin(bx) + c \cdot \cos(dx) + e \cdot \sin(fx) + g \cdot \cos(hx) \tag{12}$$

yields an almost perfect fit (with R^2 = 0.99). This can also be seen from Figure 5, which shows the observed data points as well as the regression curve based on equation (9). This result proves that quite obviously there is some regular organization in the sequential order of word length in the proverbs examined here – but although such a model would be mathematically convincing, it would hardly allow for any simple linguistic or paroemiological explanation, as far as its parameter values are concerned.[10]

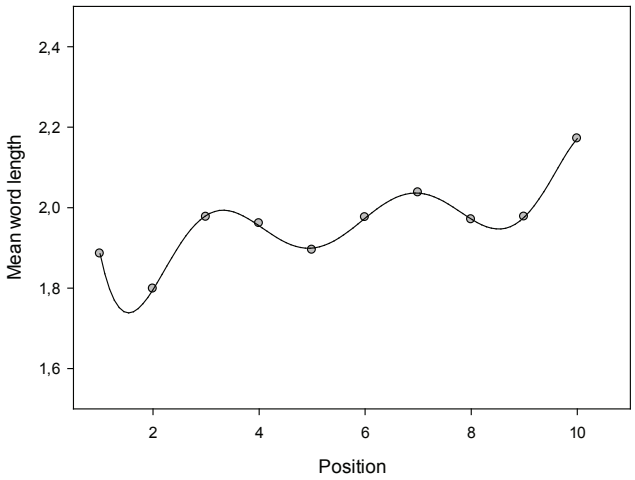


Figure 5. *Fourier series for mean word length, given separately for individual within-proverb positions.*

With this in mind, it might therefore be more promising to expand a line suggested by Krikmann, and additionally take into account sentence length again: in the two approaches discussed above, we have analyzed

[10] There is no need to mention the parameter values here. Suffice it to say, that the situation is not essentially different, if (12) is reduced by two parameters, i.e., $f(x) = k + a \cdot \sin(bx) + c \cdot \cos(dx) + e \cdot \sin(x)$, the fitting result in this case still being very good (R^2 = 0.97).

a) the dependence of word length on sentence length, paying no attention to the specific within-sentence position, and
b) the dependence of word length on within-proverb position, ignoring specific proverb length.

For both questions, we can find a model suitable to cover the underlying systematic mechanism. Thus, what is still lacking is a combination of both approaches, i.e., an analysis of average word length for each individual position within a proverb sentence, but separately for each individual sentence length. There is no need to reproduce the corresponding data in detail here, which can be found in Krikmann's (1967: 133ff) text, and which are graphically presented in Figure 6 in the same form as in that text:

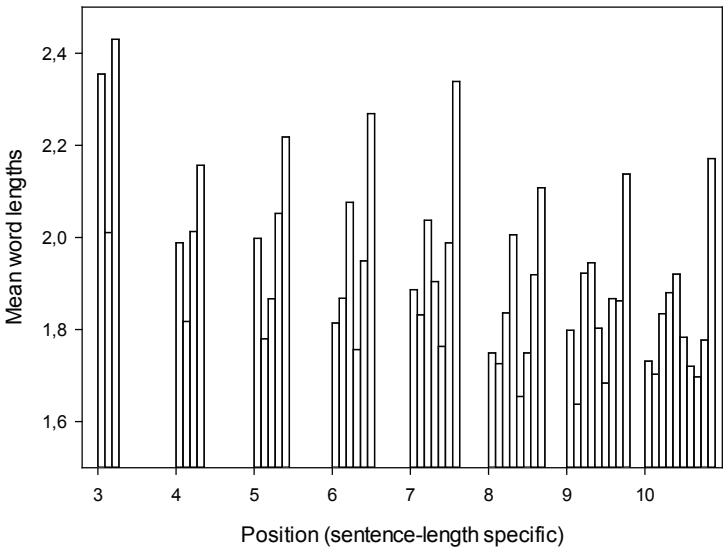


Figure 6. *Mean word length for individual within-proverb positions, given separately for individual proverb (sentence) lengths.*

Analyzing the given data situation, Krikmann (1967: 139) arrived at the conclusion[11] that there are essentially two different kinds of structure: the first type is represented by those proverbs with a sentence length

[11] A number of further observations made by Krikmann can be neglected here.

from three to five words (T_3–T_5, in Krikmann's terminology adopted here and for the remainder of this chapter), the second type by those from six to 10 words (T_6–T_{10}); accordingly, the first type is unipartite, the second is bipartite.

The shorter unipartite proverbs are, as regards the position-dependent order of word length, characterized by an initial decrease, followed by an increase; more specifically, the minimal word length in all of these cases – i.e. for T_3, T_4, and T_5 – is the second position (Pos_2), the maximum is the last position (i.e., Pos_3, Pos_4, or Pos_5, respectively).

As compared to this, the longer bipartite proverbs (T_6–T_{10}) are characterized by two periods, or cycles: the first of these two cycles behaves principally in the same way as the proverbs from T_3–T_5; as to the second cycle, proverbs of T_7, T_9, and T_{10} also behave as the first cycle, whereas the second cycles of proverbs with T_6 and T_8 are characterized by a monotonous increase.

Starting from Krikmann's observations, and attempting to expand on them, it is tempting to see if the observed tendencies and regularities follow a common tendency and can be grasped by a functional model. Searching for such a model, it seems reasonable to 'split' the curves for the longer bipartite proverbs and consider them to be composed of two separate parts; as a consequence, we are concerned with 13 curves (the curves for $T_{\geq 6}$ being composed of an a and a b part). Moreover, a relatively complex model is to be expected, if this model is to cover all curves, since despite the overall similar tendency, the degrees of decrease and increase differ for all of them.

Thus, assuming that the characteristic decreasing → increasing tendencies are the result of specific underlying processes, we can assume that this tendency, which the exception of the monotonously increasing curves at T_6 and T_8, is essentially ruled by (at least) two processes that are antagonistic by nature. This leads to the assumption that the relative rate of change is composed of a constant (a), on the one hand, and a sum (or, in case of $c < 0$, a difference) of differently weighted proportional processes, on the other; if we tentatively set these as b/x and d/x^2, we obtain the differential equation (9a) already mentioned above

$$\frac{y'}{y} = b + \frac{a}{x} + \frac{c}{x^2}, \tag{10a}$$

resulting in function (10) above.

Thus, tentatively supposing that the observed tendencies can in principle be grasped by function (10), it seems to be justified to additionally expect that at least for the shorter cycles, minimally one of the parameter values converges to 0 or 1, resulting in a simpler model. It goes without saying that, if the assumption holds, this needs not be one and the same parameter for all curves. In any case, a simplified model of (10) will inevitably be needed for the curves of T_3, T_{6a}, T_{6b} and T_{7a}, with only three words each, since otherwise the model would have more parameters than data points.

It seems reasonable, therefore, to start the analysis with the cycles composed of minimally four words. Table 6 presents the fitting results, i.e. the parameter values and the corresponding R^2 values, which in the next step can serve as some kind of benchmark for some simplified model.

Table 6. *Results of fitting function (10).*

	K	**a**	**b**	**c**	
		a_0	a_1	a_2	R^2
T3	---	---	---	---	---
T4	0.3159	-0.2650	1.7708	-2.1049	>0.9999
T5	0.8016	0.0391	0.4035	-0.8744	0.9971
T6a	---	---	---	---	---
T6b	---	---	---	---	---
T7a	---	---	---	---	---
T7b	0.7720	0.1781	0.1548	-0.7246	>0.9999
T8a	1.4021	0.1445	-0.1724	-0.0766	>0.9999
T8b	1.1934	0.0659	0.1732	-0.2608	>0.9999
T9a	0.0393	-0.9206	4.6157	-4.7446	>0.9999
T9b	1.3042	0.1225	-0.1053	-0.1998	0.8926
T10a	0.6878	-0.1525	0.9765	-1.0754	0.9888
T10b	9.8301	0.6655	-2.7136	2.3734	0.9853

As a first glance at Table 6 shows, the fitting results are very good in all cases. As can also be seen, in almost all cases (with the exception of T10b) at least one of the parameters (a, b, c) seems to converge to 0 or 1, thus indeed resulting in a simplified model; however, there

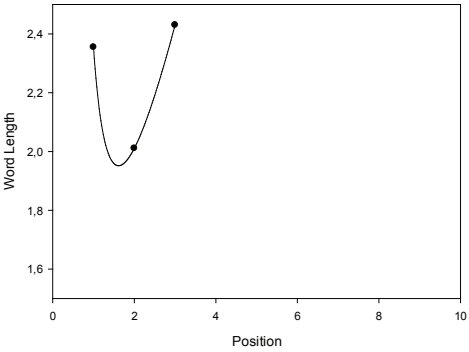
seems to be no common tendency as to the concrete parameter for the individual cycles; rather, it seems we are concerned with various kinds of 'local' modifications (or simplifications) of the general model, and it is a matter of empirical approach which of these simpler model adequately characterizes the given cycle.

Faced with the task of identifying suitable modifications, with regard to the shorter cycles, on the one hand, and the desire to find a simplified model (with fewer parameters) for the longer ones, on the other, the procedure should not, of course, result in a significant decrease of the model fit; therefore I set a benchmark of $R^2 \geq 0.99$ for all individual cycles (and $R^2 = 0.89$ for T9b, respectively).

The following tables and figures show the corresponding results. It turns out that in all cases, we are concerned either with the combination of two exponential functions, or of an exponential and a power function. The tables contain for each cycle the adequate function, along with the empirical mean values for word length at a given position, the parameter values, the discrepancy coefficient C, and the accompanying illustration, with the filled circles representing the observed values, and the function representing the theoretical curve.

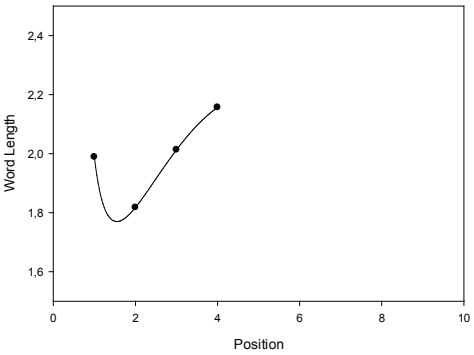
T_3: $y = K \cdot e^{ax} \cdot e^{1/x} = K \cdot e^{(ax+1/x)}$

1	1.9981	$\boldsymbol{K}$	= 0.6098
2	1.7799	$\boldsymbol{a}$	= 0.34912
3	1.8668	$\boldsymbol{R^2}$	> 0.99



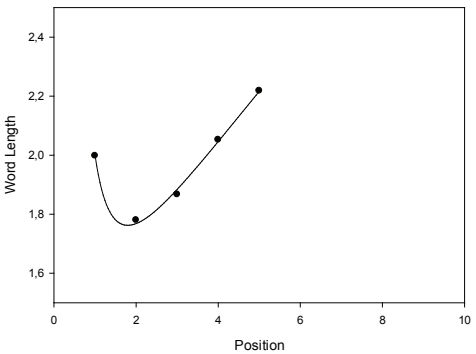
T_4: $y = K \cdot e^{ax} x^b \cdot e^{c/x}$

1	1.9981	***K***	= 0.3159
2	1.7799	***a***	= -0.2650
3	1.8668	***b***	= 1.7708
4	2.0521	***c***	= -2.1049
		$\boldsymbol{R}^2$	> 0.99



T_5: $y = K \cdot x^b \cdot e^{1/x}$

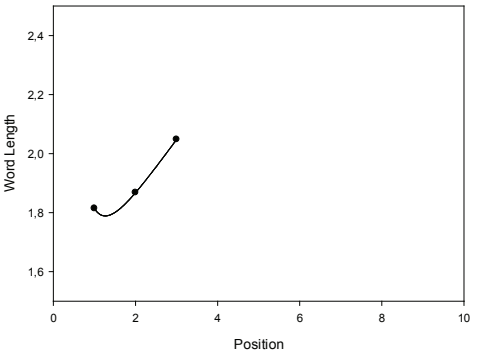
1	1.9981		
2	1.7799	***K***	= 0.7322
3	1.8668	***b***	= 0.5612
4	2.0521	$\boldsymbol{R}^2$	> 0.99
5	2.2181		



T_{6a}: $y = K \cdot x^b \cdot e^{c/x}$

1 1.8141
2 1.8676
3 2.0479

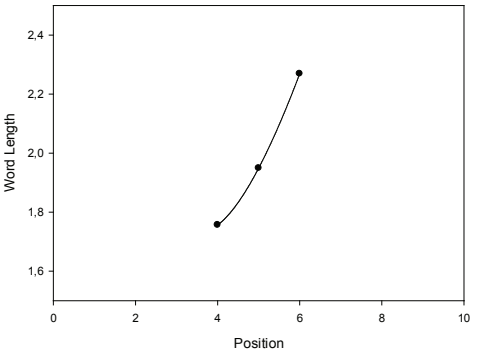
K = 0.9982
b = 0.4728
c = -0.5974
R^2 > 0.9999



T_{6b}: $y = K \cdot (x-3)^b \cdot e^{c/(x-3)}$

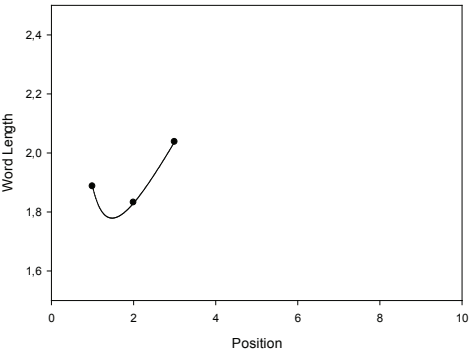
4 1.8141
5 1.8676
6 2.0479

K = 0.5955
b = 0.4728
c = -0.5974
R^2 > 0.99



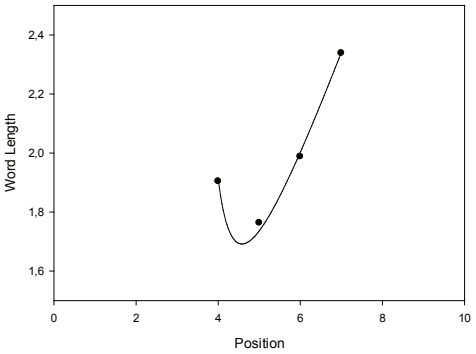
T_{7a}: $y = K \cdot x^{b} \cdot e^{1/x}$

1 1.8865 ***K*** = 0.6942
2 1.8317 ***b*** = 0.6769
3 2.0372 $\boldsymbol{R}^2$ > 0.99



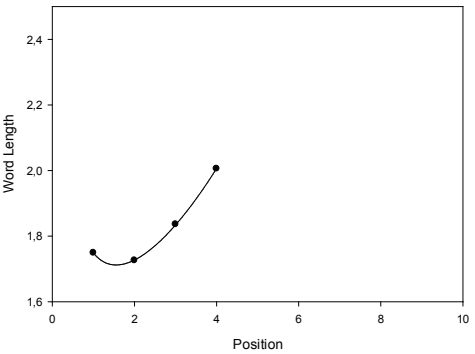
T_{7b}: $y = K \cdot (x-3) \cdot e^{c/(x-3)}$

4 1.9041
5 1.7632 ***K*** = 0.3945
6 1.9883 ***c*** = -1.5774
7 2.3386 $\boldsymbol{R}^2$ > 0.99



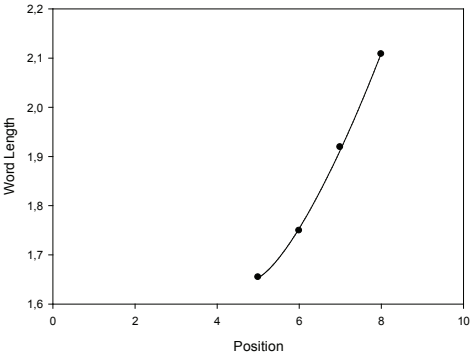
T_{8a}: $y = K \cdot e^{ax} \cdot x^{b}$

1 1.7490 ***K*** = 1.4869
2 1.7259 ***a*** = 0.1622
3 1.8359 ***b*** = -0.2519
4 2.0058 $\boldsymbol{R}^2$ > 0.99



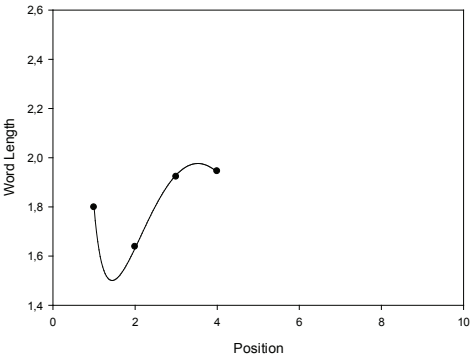
T_{8b}: $y = K \cdot (x-4)^{b} \cdot e^{a(x-4)}$

5 1.6544 ***K*** = 1.4576
6 1.7490 ***a*** = 0.1259
7 1.9189 ***b*** = -0.0964
8 2.1081 $\boldsymbol{R}^2$ > 0.99



T_{9a}: $y = K \cdot e^{x} \cdot x^{b} \cdot e^{-c/x} = K \cdot x^{b} \cdot e^{(x-c/x)}$

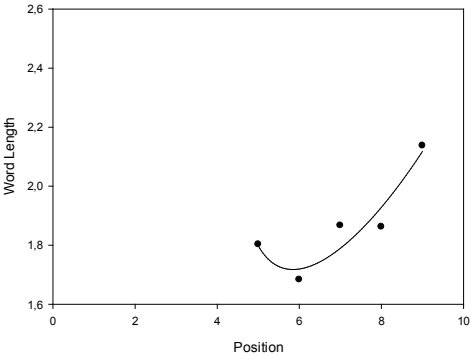
1 1.7982 ***K*** = 0.0304
2 1.6376 ***b*** = 4.9684
3 1.9220 ***c*** = -5.0820
4 1.9450 $\boldsymbol{R}^2$ > 0.99



T_{9b}: $y = K \cdot e^{ax} \cdot x^{b}$

1 1.8028
2 1.6835
3 1.8670
4 1.8624
5 2.1376

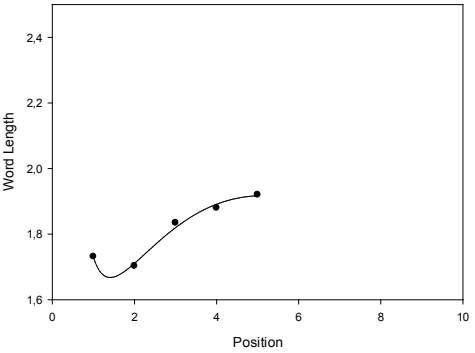
K = 1.5329
A = 0.1595
b = -0.2946
$\boldsymbol{R}^2$ = 0.89



T_{10a}: $y = K \cdot e^{ax} \cdot x^{b} \cdot e^{1/x} = K \cdot x^{b_1} \cdot e^{(ax+1/x)}$

1 1.7314
2 1.7029
3 1.8343
4 1.8800
5 1.9200

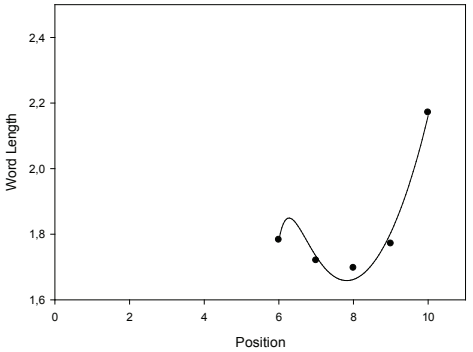
C = 0.7310
a = -0.1386
b = 0.9050
R^2 = 0.99



T_{10b}: $y = K \cdot e^{ax} \cdot x^{b} \cdot e^{c/x}$

6 1.7829
7 1.7200
8 1.6971
9 1.7714
10 2.1714

K = 10.2711
a = 0.6764
b = -2.7680
c = 2.4281
R^2 = 0.98



Conclusions

As the above analyses show, the linguistic organization of Estonian proverbs is far from being random: rather, it is characterized by regular mechanisms. In the present contribution, focusing on word length only, it can be shown that these regularities concern not only word length frequency organization, but also position-dependent specifics, as well as dependencies between sentence length and word length. Not only can all of the resulting tendencies of the underlying organizational processes be stochastically modelled, moreover, this is possible in the general framework of well-known linguistic theories. It remains open to corroborate these results on a larger database and to compare them with regard to proverbs from more languages, on the one hand, and to the general linguistic situation in Estonian, on the other.

References

Anderson, Walter 1935. *Studien zur Wortsilbenstatistik der älteren estnischen Volkslieder.* Acta et Commentationes Universitatis Tartuensis (Dorpatensis). B; 34.1 Eesti Rahvaluule arhiivi toimetused = Commentationes archivi traditionum popularium Estoniae 2, Tartu.

Antić, Gordana & Kelih, Emmerich & Grzybek, Peter 2005. Zero-syllable Words in Determining Word Length. In: Grzybek, Peter (ed.) *Contributions to the Science of Text and Language. Word Length Studies and Related Issues. Text, Speech and Language Technology*, Vol. 31. Dordrecht, NL: Springer, pp. 117–156.

Bartens, Hans-Hermann & Best, Karl-Heinz 1996. *Ural-Altaische Jahrbücher*, N.F. Vol. 14, pp. 112–128.

Grotjahn, Rüdiger & Altmann, Gabriel 1993. Modelling the Distribution of Word Length: Some Methodological Problems. In: Köhler, Reinhard & Rieger, Burghard (eds.) *Contributions to Quantitative Linguistics.* Dordrecht, NL: Kluwer Academic Publishers, pp. 141–153.

Grzybek, Peter 2000. Zur Wortlänge und ihrer Häufigkeitsverteilung in Sprichwörtern (Am Beispiel slowenischer Sprichwörter, mit einer Re-Analyze estnischer Sprichwörter). In: Palm-Meister, Christine (ed.) *Europhras 2000. Internationale Tagung zur Phraseologie vom 15.–18. Juni 2000 in Aske / Schweden*. Tübingen: Stauffenburg, pp. 161–171.

Grzybek, Peter 2006. History and Methodology of Word Length Studies. The State of the Art. In: Grzybek, Peter (ed.) *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Dordrecht, NL: Springer, pp. 15–90. [Text, Speech and Language Technology; 31.]

Grzybek, Peter 2014. Word Length. In: Taylor, John (ed.) *Handbook of the Word.* Oxford: Oxford University Press. [Forthcoming.]

Krikmann, Arvo 1967. Keelestatistikat Eesti vanasõnadest. [Linguostatistics of Estonian Proverbs.] *Emakeele Seltsi aastaraamat* [Yearbook of the Society for the Estonian Language], Vol. 13, pp. 127–154.

Krikmann, Arvo & Sarv, Ingrid 1996. The Tartu Research Group of Paremiology. *Folklore: Electronic Journal of Folklore*, Vol. 2, pp. 87–115. http://www.folklore.ee/folklore/vol2/mieder.htm, last accessed on December 10, 2013.

Leino, Pentti 1968. Mitteilung. *Proverbium*, Vol. 11, p. 302.

Wimmer, Gejza & Altmann, Gabriel 1996. The Theory of Word Length: Some Results and Generalizations. *Glottometrika*, Vol. 15, pp. 112–133.

Wimmer, Gejza & Altmann, Gabriel 1999. *Thesaurus of Univariate Discrete Probability Distributions.* Essen: Stamm-Verlag.

Wimmer, Gejza & Altmann, Gabriel 2005. Unified Derivation of Some Linguistic Laws. In: Köhler, Reinhard & Altmann, Gabriel & Piotrowski, Rajmund G. (eds.) *Quantitative Linguistics: An International Handbook. Quantitative Linguistik: Ein Internationales Handbuch.* Berlin: de Gruyter, pp. 791–807.

Wimmer, Gejza & Altmann, Gabriel 2006. Towards a Unified Derivation of Some Linguistic Laws. In: Grzybek, Peter (ed.) *Contributions to the Science of Text and Language: Word Length Studies and Related Issues.* Dordrecht, NL, pp. 329–337.

Wimmer, Gejza & Köhler, Reinhard & Grotjahn, Rüdiger & Altmann, Gabriel 1994. Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, Vol. 1, pp. 98–106.