Journal of Quantitative Linguistics
2003, Vol. 00, No. 0, pp. 000–000

0929-6174/03/0000-000$16.00

# A Quantitative Approach to Lexical Structure of Proverbs

Peter Grzybek
Institut für Slawistik, Universität Graz, Graz, Austria

## INTRODUCTION

As has been repeatedly pointed out elsewhere, the linguistic structure of proverbs has hardly ever been seriously studied with regard to underlying regularities. Basically, proverb research has hardly ever transgressed the level of symptomatic descriptions and not yet reached a systematic level (Grzybek 2000a, 2000b, 2001, 2002).

Quite logically, first attempts at filling the evident gaps in this field of research have concentrated on rather special problems, such as on the question if sentence length or word length of proverbs can be systematically described. In this respect, special attention has been paid to the development of adequate methods which go beyond traditional approaches, confining themselves to the tabulation of absolute or relative frequencies, their means, or the presentation of simple graphs. Specifically, the seemingly trivial questions have been asked, how often words or sentences of a given length occur in proverbs, and if these frequencies can be modelled by way of mathematical and/or statistical devices. Irrespective of the seeming simplicity of these questions, a number of successive assumptions come into play which make the underlying complexity of the whole problem evident; it is assumed that:

- the frequency with which proverbs of a given length (or words as verbal constituents of proverbs) occur in a proverb corpus is not accidentally (chaotically) organized, but follows particular regularities;
- it is possible to describe and to formalize these regularities;

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

- the formalization of these regularities allows for cross-references to general empirical observations and theoretical assumptions in quantitative linguistics;
- these cross-references result in assumptions as to the specific linguistic organization of proverbs.

In order to pursue the questions delineated above, it seemed reasonable to take into consideration the general theoretical framework outlined by Wimmer, Köhler, Grotjahn, and Altmann (1994), or Wimmer and Altmann (1996), against the background of a general synergetic approach. In the following ruminations, as well, it will turn out to be useful to rely on these concepts, although the questions to be asked will slightly differ from the foregoing. The question which shall be focused on in this article will concentrate on nothing more (and nothing less) but the question if the lexical repertory of a given proverb corpus is characterized by particular regularities as regards its lexical frequency structure. In detail, the question to be studied is, if the frequencies with which the single words occur within a given proverb corpus, can be theoretically modelled. In other words: whereas previous studies concentrated on the length of the units to be studied (and on the frequency of their occurrence), the present study focuses on the lexical frequency structure itself.

## MATERIAL

Our material to be studied is the collection of Slovenian proverbs Pregovori, prilike in reki by Kocbek (1887). This first comprehensive collection of Slovenian proverbs can be regarded to be the ground work of Slovenian paremiography, at best, which also represents the basis for later collections, as those by Kocbek and Šašelj (1934), Bojc (1974, 1980, 1987) or Prek (1972, 1974, 1982, 1986, 1996).

The Kocbek collection contains 2.429 proverbial sentences, consisting of 15.467 word form tokens, based on 4.638 word form types. The observation that 2.887 word forms occur exactly once – i.e., the so-called ‘hapax legomena’ (thus representing 18.66% of the word form tokens and 62.25% of the word form types) – leads to the question of a systematic study of lexical frequencies.

Given that each of the $n$ word forms occurs with a minimal frequency of $I = 1$ and a maximum of $i = m$ occurrences, we can state that in our case the

Table 1. Rank Frequency List of the 10 Most Frequent Word Forms.

| $r$ | | $f_r$ | $p_r$ |
|---|---|---|---|
| 1 | ne | 487 | 0.0315 |
| 2 | je | 455 | 0.0294 |
| 3 | se | 381 | 0.0246 |
| 4 | kdor | 264 | 0.0171 |
| 5 | v | 241 | 0.0156 |
| 6 | na | 195 | 0.0126 |
| 7 | pa | 125 | 0.0081 |
| 8 | za | 109 | 0.0070 |
| 9 | in | 106 | 0.0069 |
| 10 | ima | 101 | 0.0065 |
| ... | ... | ... | ... |

amount of word forms is $n = 15.467$, and $1 \leq I \leq 487$; and terming the concrete number of occurrences of each word form $f_i$, we can state that in our case the maximum of $f_i$, is 2.887 for $i = 1$, and the minimum is $f_i = 1$ for $i = 487$. We thus obtain

$\sum_{i=1}^{m} f_i = 4.638$ for the amount of word form types, and $\sum_{i=1}^{m} i \cdot j_i = 15.467$ for the amount of word form tokens.

As was indicated above, the present article will not deal with the question which concrete word forms occur with which frequency; still, it might be interesting to take a look at the most frequent word forms. Table 1 represents a rank frequency table for the ten most frequent word forms; for each rank $r$ ($r = 1, 2, \ldots, m$), it contains the concrete word form, and in addition to this, the absolute ($f_r$) and relative ($p_r$) frequency of these word forms. In our case, the ten most frequent word forms are those with a frequency of $f_r > 100$.

Interestingly enough, eight of the ten most frequent word forms are also contained in the "Top 10" list of the electronic corpus of Slovenian literary texts (CORTES)[1] – in that corpus, only 'kdor' [when] and 'ima' [s/he has] have a clearly lower frequency (rank 330, and rank 121, respectively). This observation allows for the hypothesis, that our proverbial material might well

[1] Here, the material basis is a word frequency list of the 1.000 most frequent Slovenian words from the CORTES corpus, compiled by Primož Jakopin. At the time of its analysis (December, 1999), this corpus consisted of 112 literary (mainly prose) texts from the 19th and 20th centuries. The texts were written by 41 authors; 98 of the texts were original Slovene texts, 14 were Slovenian translations from other languages.

be characterized by specifics which, on the one hand, are coined by general linguistic characteristics, and which, on the other hand, are characterized by specifics of its own. Pursuing this assumption, three interrelated questions shall be dealt with in the following analyses:

1. *Rank frequency distribution*: How often do the most frequent, the second most frequent, etc. word forms occur? Is there a specific relation between the individual frequency classes? Is it possible to interpret this relation in terms of a functional relationship $P_x = g(x)P_{x-1}$, resulting in a dynamic system?
2. *Frequency spectrum*: How many word forms occur exactly $1, 2, 3, \ldots, n$ times in our corpus?
3. *Lexical coverage*: Which percentual part of the whole lexical inventory is covered by the most frequent, the two, three, etc. most frequent word forms?

It is obvious that these three questions are closely related to each other; still, let us naively start with the first of them. We are concerned here with a problem, which is well known since more than a half century, and which is initially related to the name and the work of George Kingsley Zipf.

## ZIPF AND MANDELBROT

Since Zipf's ideas should be well known in quantitative linguistics, they shall only briefly be called to mind, here. In his book *The Psycho-Biology of Language. An Introduction to Dynamic Philology*, published in 1935, Zipf provided the foundations for a first concept of word frequency; in it, he argued in favor of the notion that the frequency of words in texts is not randomly, but regularly organized. In detail, he postulated a relation between the frequency of a word and the number of words which display this frequency. According to his observations, there are only a few words in a text, which occur relatively often, and there are many words which occur only rarely. Zipf combined this observation – i.e., the assumption of decreasing variability going along with an increasing frequency – with the hypothesis that we are concerned here with a law-like relation which he tried to describe by a relatively simple mathematical equation:

$$a \cdot b^2 = k \tag{1}$$

Here, *a* corresponds to the number of words with a given frequency, *b* corresponds to the number of occurrences, and *k* is a particular constant, characteristic for the given text. According to this formula, the product of the squared frequency of a given word and the sum of its occurrences would thus turn out to be constant.

In a later attempt to formalize his observations, Zipf (1949) took into consideration the absolute frequency ($f$) of the elements, as he did before, but, in addition to this, he now introduced the rank ($r$) a word has within a given text (corpus). This led him to the following equation:

$$r \cdot f = k \tag{2}$$

According to this approach, the product of the absolute frequency of a word and its rank turns out to be a constant.

Both approaches led to convincing results, with one exception reported by Zipf (1935, p. 43), namely the most frequent and the most rare words. Due to this reason, Zipf's concept has been repeatedly modified, extended or transformed into more general models. Sometimes, these models give rise to the impression of a scholarly discipline in its own right, for the understanding of which a diploma in mathematics is necessary (cf., e.g., Baayen, 2001; Guiter & Arapov, 1982, etc.). As to language and linguistic texts, one of the best known and most important generalization of Zipf's formula is the one by Mandelbrot (1953, 1954); he picked up the thread of Zipf by elaborating on the observation that the regularity described by Zipf is valid for the "intermediate" area of a text's vocabulary, but not for the extreme (i.e., the most frequent and the most rare) occurrences. Mandelbrot started from the simple Zipf formula

$$r \cdot f = C \tag{2$'$}$$

which, after a slight transformation, can be read as

$$f = \frac{C}{r} \tag{2a}$$

In a next step, absolute frequency can be understood to be a function of the rank $r$ and the constant $C$; additionally defining rank as a variable $x$, one obtains

$$f(x) = \frac{C}{x} \tag{2b}$$

This function can easily be transformed into a probability function, allowing to calculate the theoretical frequency $P_x$ for each $x$. For this purpose, Equation (2b) has to be truncated to the right (because the harmonic series does not converge – i.e., the range of definition now is $x = 1, 2, 3, \ldots, n$). In this process, $C$ turns out to be a norming constant, defined as $C^{-1} = \psi(n+1) - \psi(1)$, where $\psi$ is the digamma function (i.e., the logarithmized gamma function):

$$P_x = \frac{1}{x[\psi(n+1) - \psi(1)]} \quad x = 1, 2, 3, \ldots, n \tag{2b$'$}$$

This distribution is called Estoup distribution (Wimmer & Altmann, 1999); a generalization of Equation (2b) is

$$P_x = \frac{x^{-a}}{T(a)} \quad x = 1, 2, 3, \ldots, n; \; a \in \Re; \qquad T(a) = \sum_{j=1}^{n} j^{-a} \tag{2c}$$

This probability mass function is usually called the "classical" Zipf distribution (or right-truncated Zeta distribution), which differs from Zipf's original assumptions in two regards: (a) the norming constant $C$ cannot be presented in closed form, (b) the exponent $a$ is $\neq 1$, whereas the original form postulated $a = 1$. As compared to (2c), the distribution contemporarily called Zipf–Mandelbrot distribution contains the additional parameter $b$, resulting in

$$P(x) = \frac{C}{(b+x)^a} \quad x = 1, 2, 3, \ldots, n \tag{3}$$

A closer look at Equation (3) shows that Zipf's original formula (2b) turns out to be a special case of the Zipf–Mandelbrot Equation (3) in form of a probability function (2b$'$), namely, if $a = 1$ und $b = 0$. Hereby, the constant $C$ of probability function (3) is a norming constant, which can be estimated from parameters $a$ and $b$:

$$C^{-1} = \sum_{i=1}^{n} (b+i)^a \tag{3$'$}$$

The distribution model thus obtained has proven to be adequate for modelling word frequencies, primarily of (longer) texts and word frequency lists on the basis of dictionaries or text corpora. In the following analyses, it shall be tested if Zipf's theoretical approach can also be applied to describe the lexical structure of proverbs.

This methodological transfer to proverbial material is an innovation and far from self-evident; more likely than not, this transfer will even seem questionable to many a (quantitative) linguist, since a proverb collection is neither a homogeneous text – rather, each proverb can be regarded to be a closed text in itself – nor a lexicon based on a text corpus in the ordinary sense of this word. Yet, one might argue in favor of the notion that a proverb collection can be regarded as a proverb lexicon in the strict sense of this word, the entries of which have to be seen on a sentence level, not a lexical level. If, therefore, the application of Zipf's approach to our proverb material should turn out to be successful, this would be a strong argument in favor of the assumption that the lexical structure of proverbs is systematically organized, and that the language of proverbs is characterized by strict regularities.

## ANALYSES

As was pointed out above, the following analyses will concentrate on three topics:

*lexical rank frequency distribution*,
*lexical frequency spectrum*,
*lexical coverage.*

It goes without saying that it cannot be sufficient to simply present the empirical findings; rather, it will be important to test how the data may be described on the basis of theoretical models. Furthermore, attention should be paid to the fact that these three questions are closely interrelated, the more so as a frequency distribution can be mathematically transformed both into a frequency spectrum and a description of text coverage. In the given context, the individual questions shall be focused, however, since, in this first step, the overall objective is to make evident the relevance of Zipf–Mandelbrot's law for proverb research. Let us start, therefore, with the rank frequency distribution, which represents something like a starting point in Zipf's concept, as well.

### Rank Distribution

The question as to the absolute frequency ($f_i$) of the most frequent, the second most frequent, etc. word form ($i = 1, 2, \ldots, n$) basically contains Zipf's search for a rank frequency distribution. Since, in our case, we are concerned

Table 2. Rank Frequency Distribution of the Word Forms.

| $i$ | $f_i$ | $NP_i$ | $i$ | $f_i$ | $NP_i$ | $i$ | $f_i$ | $NP_i$ | $i$ | $f_i$ | $NP_i$ | $i$ | $f_i$ | $NP_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 487 | 523.50 | 11 | 100 | 121.80 | 21 | 72 | 71.34 | 31 | 55 | 51.04 | 41 | 45 | 39.98 |
| 2 | 455 | 386.47 | 12 | 94 | 113.57 | 22 | 70 | 68.57 | 32 | 53 | 49.65 | 42 | 43 | 39.14 |
| 3 | 381 | 307.89 | 13 | 92 | 106.42 | 23 | 66 | 66.02 | 33 | 53 | 48.34 | 43 | 42 | 38.34 |
| 4 | 264 | 256.71 | 14 | 92 | 100.15 | 24 | 64 | 63.66 | 34 | 53 | 47.10 | 44 | 42 | 37.57 |
| 5 | 241 | 220.62 | 15 | 86 | 94.62 | 25 | 63 | 61.47 | 35 | 52 | 45.92 | 45 | 40 | 36.83 |
| 6 | 195 | 193.75 | 16 | 85 | 89.68 | 26 | 63 | 59.43 | 36 | 49 | 44.80 | 46 | 38 | 36.12 |
| 7 | 125 | 172.93 | 17 | 84 | 85.26 | 27 | 62 | 57.53 | 37 | 49 | 43.74 | 47 | 38 | 35.44 |
| 8 | 109 | 156.31 | 18 | 81 | 81.27 | 28 | 62 | 55.75 | 38 | 47 | 42.73 | 48 | 37 | 34.79 |
| 9 | 106 | 142.72 | 19 | 80 | 77.65 | 29 | 59 | 54.08 | 39 | 46 | 41.77 | 49 | 35 | 34.16 |
| 10 | 101 | 131.39 | 20 | 73 | 74.36 | 30 | 59 | 52.52 | 40 | 45 | 40.85 | 50 | 34 | 33.56 |

with no less than 15.467 rank positions, the whole sample was divided at the median; we are dealing, therefore, with the upper 50 percent of a right-truncated distribution, which cover 241 ranks.

Table 2 presents the data for the first 50 of these 241 ranks: in the first column, the rank ($i$) can be found; in the second column, the absolute frequency ($f_i$), in the third column, the theoretical values ($NP_i$), which result from fitting the Zipf–Mandelbrot distribution to our data. Usually, the goodness of fit is tested by the $\chi^2$-goodness-of-fit test. Since, for large samples (characteristic in linguistic studies), this test soon displays significant results, one usually calculates the discrepancy coefficient $C = \chi^2/N$ instead. This coefficient is regarded as in index of a good fit for $C < 0.02$, as an index of an excellent fit for $C < 0.01$.

In our case, fitting the Zipf–Mandelbrot distribution results in an excellent fit ($a = 0.91$, $b = 1.53$, $n = 241$; $\chi^2 = 103.66$, $FG = 237$, $P > 0.99$); obvious deviations are be observed in the range of ranks 7 through 18 (cf. Fig. 1).

Figure 1 demonstrates the good fit of the Zipf–Mandelbrot distribution; for reasons of perspicuity, it is also confined to the first 50 ranks.

## Frequency Spectrum

Basically, the study of frequency spectra corresponds to Zipf's earlier ruminations, in which the rank of occurrence did not play a crucial role. Meanwhile, however, it has repeatedly been shown that rank frequency and frequency spectrum can be mutually transformed (cf. Chitashvili & Baayen, 1993; Zörnig & Boroda, 1992). In detail, the question at stake is, how many word forms there are ($f_i$), which occur $i = 1, 2, 3, \ldots, m$ times; given the raw

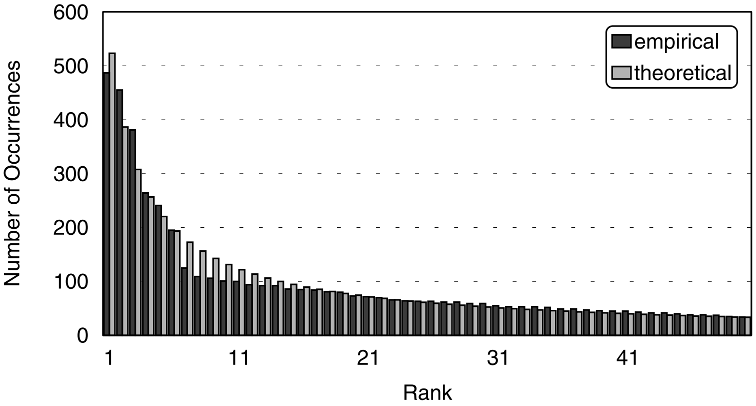

Fig. 1. Rank frequency distribution of the word forms.

Table 3. Absolute Frequency of the Word Forms.

| $i$ | $f_i$ | $NP_{(i)}$ | $i$ | $f_i$ | $NP_{(i)}$ | $i$ | $f_i$ | $NP_{(i)}$ | $i$ | $f_i$ | $NP_{(i)}$ | $i$ | $f_i$ | $NP_{(i)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2887 | 2772.57 | 11 | 19 | 18.61 | 21 | 4 | 3.90 | 31 | 1 | 1.50 | 41 | 0 | 0.75 |
| 2 | 732 | 807.63 | 12 | 13 | 15.12 | 22 | 0 | 3.48 | 32 | 1 | 1.39 | 42 | 2 | 0.71 |
| 3 | 314 | 354.59 | 13 | 14 | 12.48 | 23 | 3 | 3.13 | 33 | 2 | 1.29 | 43 | 1 | 0.67 |
| 4 | 171 | 191.00 | 14 | 11 | 10.44 | 24 | 2 | 2.82 | 34 | 3 | 1.20 | 44 | 0 | 0.63 |
| 5 | 119 | 116.32 | 15 | 10 | 8.84 | 25 | 1 | 2.55 | 35 | 1 | 1.11 | 45 | 2 | 0.60 |
| 6 | 85 | 76.90 | 16 | 8 | 7.56 | 26 | 3 | 2.31 | 36 | 0 | 1.04 | 46 | 1 | 0.57 |
| 7 | 54 | 53.91 | 17 | 10 | 6.53 | 27 | 2 | 2.11 | 37 | 1 | 0.97 | 47 | 1 | 0.54 |
| 8 | 41 | 39.50 | 18 | 9 | 5.68 | 28 | 2 | 1.93 | 38 | 2 | 0.91 | 48 | 0 | 0.51 |
| 9 | 29 | 29.96 | 19 | 7 | 4.98 | 29 | 4 | 1.77 | 39 | 0 | 0.85 | 49 | 2 | 0.48 |
| 10 | 19 | 23.35 | 20 | 8 | 4.40 | 30 | 1 | 1.63 | 40 | 1 | 0.80 | 50 | 0 | 0.46 |

data, it can be tested if the Zipf–Mandelbrot model in this case, too, turns out to be adequate. Table 3 contains the relevant data from our proverbial material.

As can be seen from Table 3, there are 2.887 word forms, which occur exactly once, 732 word forms, which occur twice, etc. We are thus concerned here with the most rarely occurring word forms; it can easily be seen that the ten rarest word forms represent almost 96%, the twenty rarest word forms almost 98% of all lexical occurrences. Figure 2 represents the data of those word forms which occur 1 through 50 times (on the whole, only 35 word forms have a frequency higher than this); Figure 2 illustrates the fit of the Zipf–Mandelbrot distribution to our data, which has to be regarded as excellent ($a = 2.51$, $b = 0.57$, $n = 50$; $\chi^2 = 43.99$, $FG = 38$, $P = 0.23$).

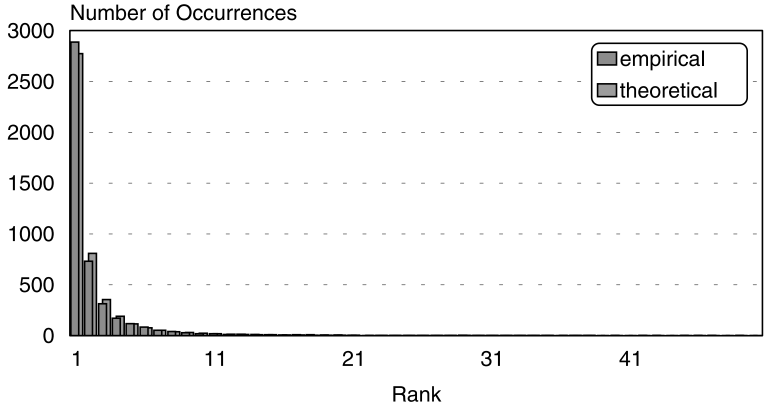

Fig. 2. Number of word forms with a given frequency.

**Lexical Coverage**

The question of lexical coverage, too, stands in close correlation to the two questions above, and text coverage, too, can be mathematically derived from a rank frequency distribution. Lexical coverage aims at the (relative) percentage, covered by the most frequent, the two most frequent, etc. word forms: As can be seen from Table 4, 3.15% of all lexical occurrences are covered by the most frequent word form (‘ne’), the two most frequent word forms taken together cover 6.09%, the ten most frequent taken together 15.93%. In quantitative linguistics, one usually speaks of “text coverage”, in this context; since we are not concerned with a homogeneous text, however, the term ‘lexical coverage’ shall be preferred in our case.

Table 4 presents the data for the first 50 positions: the first column (i) contains the ranks, the second and third row contain the cumulative absolute ($f_{cum}$) and relative ($p_{cum}$) frequencies.

In illustrating the progressively increasing process of text coverage, vocabulary-oriented studies usually have transformed the cumulative distribution function into a continuous distribution; thus, the discrete distribution has obtained the form of a continuous one. Figure 3 follows this tradition, by presenting the first 30 positions of our proverbial material, by which ca. 25% of all lexical occurrences are covered.

In order to theoretically model this trend not only for the first 30 positions, but for the whole lexical inventory, we shall try to fit a non-linear regression

Table 4. Cumulative Frequencies of the Word Forms ($N = 15467$).

| $i$ | $f_{cum}$ | $p_{cum}$ | $i$ | $f_{cum}$ | $p_{cum}$ | $i$ | $f_{cum}$ | $p_{cum}$ | $i$ | $f_{cum}$ | $p_{cum}$ | $i$ | $f_{cum}$ | $p_{cum}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 487 | 0.0315 | 11 | 2564 | 0.1658 | 21 | 3403 | 0.2200 | 31 | 4026 | 0.2603 | 41 | 4518 | 0.2921 |
| 2 | 942 | 0.0609 | 12 | 2658 | 0.1718 | 22 | 3473 | 0.2245 | 32 | 4079 | 0.2637 | 42 | 4561 | 0.2949 |
| 3 | 1323 | 0.0855 | 13 | 2750 | 0.1778 | 23 | 3539 | 0.2288 | 33 | 4132 | 0.2671 | 43 | 4603 | 0.2976 |
| 4 | 1587 | 0.1026 | 14 | 2842 | 0.1837 | 24 | 3603 | 0.2329 | 34 | 4185 | 0.2706 | 44 | 4645 | 0.3003 |
| 5 | 1828 | 0.1182 | 15 | 2928 | 0.1893 | 25 | 3666 | 0.2370 | 35 | 4237 | 0.2739 | 45 | 4685 | 0.3029 |
| 6 | 2023 | 0.1308 | 16 | 3013 | 0.1948 | 26 | 3729 | 0.2411 | 36 | 4286 | 0.2771 | 46 | 4723 | 0.3054 |
| 7 | 2148 | 0.1389 | 17 | 3097 | 0.2002 | 27 | 3791 | 0.2451 | 37 | 4335 | 0.2803 | 47 | 4761 | 0.3078 |
| 8 | 2257 | 0.1459 | 18 | 3178 | 0.2055 | 28 | 3853 | 0.2491 | 38 | 4382 | 0.2833 | 48 | 4798 | 0.3102 |
| 9 | 2363 | 0.1528 | 19 | 3258 | 0.2106 | 29 | 3912 | 0.2529 | 39 | 4428 | 0.2863 | 49 | 4833 | 0.3125 |
| 10 | 2464 | 0.1593 | 20 | 3331 | 0.2154 | 30 | 3971 | 0.2567 | 40 | 4473 | 0.2892 | 50 | 4867 | 0.3147 |

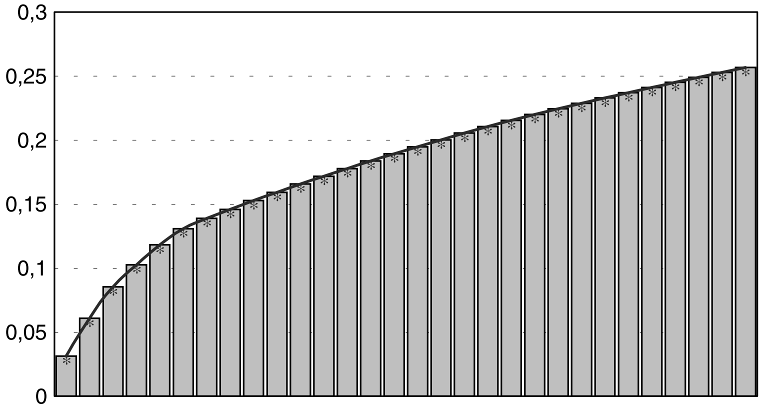

Fig. 3. Cumulative occurrences of the word forms.

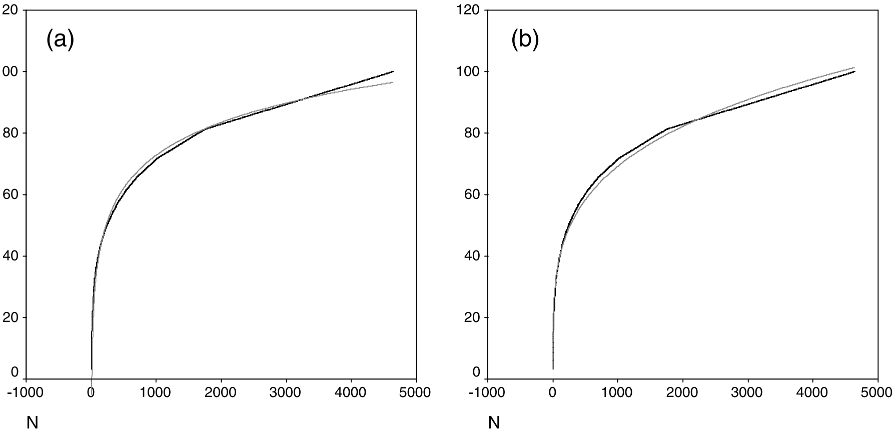

Fig. 4. (a) Logarithmic model. (b) Power model.

model to our data. Figure 4a and 4b present the result of fitting to adequate regression models:

Figure 4a shows the convincing result ($R^2 = 0.985$) of fitting a logarithmic model: $y = a + b \cdot \ln(x)$.
Figure 4b shows the result of a power model, which also turns out to be extremely adequate ($R^2 = 0.975$): $y = ax^b$.

As a result, both models lead to almost corresponding results. From a mathematical point of view, this fact is not really surprising: regarding the function in question to be a cumulative frequency function – i.e., $F(x) = P(X < x)$ – the first derivation of $F(x)$ displays the similarity of both functions:

1. For the power function we obtain: $y = a \cdot x^b \Rightarrow F'(x) = y' = a \cdot b^{(b-1)}$; after re-parametrization this leads to: $y' = A \cdot x^{(-c)}$.
2. And for the logarithmic function we obtain: $y = a + b \cdot \ln(x) \Rightarrow F'(x) = y' = bx^{(-1)}$.

Obviously, both functions differ only in their exponents. And, more importantly, both functions can now be detected to be a special case of the Zipf–Mandelbrot distribution ($3'$):

$$P_x = \frac{C}{(b+x)^a} \Rightarrow y = A \cdot (b+x)^m \tag{3''}$$

As can easily be seen, for $b = 0$, it is exactly the power function $y = A \cdot x^m$ which is obtained. Without a doubt, the derivation of the (cumulative) distribution function from the Zipf–Mandelbrot distribution is of utmost importance, since at this point, the theoretical circle of argumentation can be closed. This rather mathematically motivated question, however, would definitely go beyond the boundaries of this article and shall be reserved for a separate analysis (cf. Antić, Grzybek, & Stadlober, 2002).

## CONCLUSION

As should have become clear from the foregoing considerations and analyses, the lexical repertory of a traditional proverb collection is by far not chaotically organized, but follows particular rules. Obviously, we are concerned here with exactly the same regularities characterizing homogeneous texts and text corpora. The fact that these regularities may be successfully applied to proverbial material, is a new finding; to understand the underlying mathematical contexts will be important not only for paremiology, but also for the general study of texts.

## ACKNOWLEDGEMENT

I am sincerely grateful to G. Altmann (Lüdenscheid) and E. Stadlober (Graz) for their helpful comments and corrections.

## REFERENCES

Antić, G., Grzybek, P., & Stadlober, E. (2002). Lexikalische Vorkommenshäufigkeit und Textdeckung – Theoretische Überlegungen. In: *Glottometrics 4* (in press).

Baayen, H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Bojc, E. (1974). Pregovori in reki na Slovenskem. Državna založba Slovenije [$^2$1980, $^3$1987]

Chitashvili, R., & Baayen, H. (1993). Word frequency distributions. In L. Hřebíĉek & G. Altmann (Eds.), *Quantative text analysis* (pp. 54–135). Trier: WVT.

Grzybek, P. (2000a). Zum Status der Untersuchung von Satzlängen in der Sprichwortforschung – Methodologische Vor-Bemerkungen. In *Слово во времени и пространстве К 60-летию профессора В.М. Мокиенко* (pp. 430–457). Sankt Petersburg: Folio-Press.

Grzybek, P. (2000b). Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter. *Anzeiger für slavische Philologie*, *27*, 87–108 (1999) [2000].

Grzybek, P. (2001). Zur Satz- und Teilsatzlänge zweigliedriger Sprichwörter. In L. Uhlířová, G. Wimmer, G. Altmann, & R. Köhler (Eds.), *Text as a linguistic paradigm: Levels, constituents, constructs* (pp. 64–75). Trier: WVT.

Grzybek, P. (2002). Zur Wortlänge und ihrer Häufigkeitsverteilung in Sprichwörtern (Am Beispiel slowenischer Sprichwörter, mit einer Re-Analyse estnischer Sprichwörter). In Chr. Palm (Ed.), *Europhras 2000*. Tübingen: Narr (in press).

Guiter, H., & Arapov, M.V. (1982). *Studies on Zipf's Law.* Bochum: Brockmeyer. (= Quantitative Linguistics; vol. 16).

Jackson, W. (Ed.). (1953). *Communication theory.* London: Butterworth.

Kocbek, F. (1887). *Pregovori, prilike in reki*. Celje: Trstenjak.

Kocbek, F., & Šašelj, I. (1934). *Slovenski pregovori, reki in prilike*. Celje: Družba Sv. Mohorja.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In W. Jackson (Ed.), *Communication theory* (pp. 486–502). London: Butterworth.

Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word 10*, 1–27.

Prek, St. (1972). Ljudska modrost. Pregovori, domislice in reki. Maribor [Ljubljana, $^2$1974, $^3$1982, $^4$1986, $^5$1996].

Wimmer, G., & Altmann, G. (1996). The theory of word length distribution: Some results and generalizations. In P. Schmidt (Ed.), *Glottometrika* (Vol. 15, pp. 112–133). Trier: WVT.

Wimmer, G., & Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, *1*, 98–106.

Zipf, G. (1935). *The psycho-biology of language: An introduction to dynamic philology.* Cambridge: MIT Press (1968).

Zipf, G. (1949). *Human Behavior and the principle of least effort. An introduction to human ecology.* New York, London: Hafner (1965).

Zörnig, P., & Boroda, M. (1992). The Zipf–Mandelbrot law and the interdependencies of the frequency structure and frequency distribution in coherent texts. In B. Rieger (Ed.), *Glottometrika* (Vol. 13, pp. 205–218). Bochum: Brockmeyer.