On the law of Zipf-Mandelbrot for multi-world phrases

EGGHE, Leo (1999) On the law of Zipf-Mandelbrot for multi-world phrases. In: Journal of the American Society for Information Science, 50(3). p. 233-241.

DOI: 10.1002/(SICI)1097-4571(1999)50:3<233::AID-ASI6>3.0.CO;2-8
Handle: http://hdl.handle.net/1942/7405

ON THE LAW OF ZIPF-MANDELBROT FOR MULTI-WORD PHRASES

by

L. EGGHE

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail: legghe@luc.ac.be

ABSTRACT

The paper studies the probabilities of the occurrence of m - word phrases (m=2,3,...) in relation with the probabilities of occurrence of the single words. It is well-known that, in the latter case, the law of Zipf is valid (i.e. a power law). We prove that in the case of m - word phrases ($m \geq 2$) this is not the case. We present two independent proofs of this.
We furthermore show that - in case we want to approximate the found distribution by Zipf's law - we obtain exponents β_m in this power law for which the sequence $(\beta_m)_{m \in \mathbb{N}}$ is strictly decreasing. This explains experimental findings of Smith and Devine, Hilberg and Meyer.

---

[1]Permanent address

Acknowledgement.
The author is grateful to Prof. Dr. R. Rousseau for interesting discussions on the topic of this paper.

Our results should be compared with a heuristic finding of Rousseau who states that the law of Zipf-Mandelbrot is valid for multi-word phrases. He, however, uses other - less classical - assumptions than we do.

I. Introduction

Zipf's law (Zipf (1949)) states that if words in a text are ordered in decreasing order of occurrence in this text, then the product of the rank r of a word and the number of times it occurs $f(r)$, is a constant of that text.

$$rf(r) = c \quad . \tag{1}$$

So

$$f(r) = \frac{c}{r} \tag{2}$$

Dividing by the total number of words in the text yields the probability $P(r)$ that the word with rank r occurs:

$$P(r) = \frac{C}{r} \quad , \tag{3}$$

were C is a constant. More generally one can state

$$P(r) = \frac{C}{r^{\beta}} \tag{4}$$

where $\beta>0$, see Egghe and Rousseau (1990). In the same notation, the law of Mandelbrot states

$$P(r) = \frac{A}{(1+Br)^{\beta}} \quad . \tag{5}$$

Again A and B are constants.

This law can be found in Mandelbrot (1954,1977a, b), see also Egghe and Rousseau (1990). It is clear that laws (4) and (5) are asymptotically the same, i.e. when ranks are high:

$$\frac{A}{(1+Br)^{\beta}} = \frac{A}{\left(\left(\frac{1}{r}+B\right)r\right)^{\beta}}$$

$$\approx \frac{A}{B^{\beta}r^{\beta}}$$

$$= \frac{C}{r^{\beta}},$$

where $C=\frac{A}{B^{\beta}}$. The main part of this paper is only dealing with high ranks r and therefore we will refer to these laws as the Zipf - Mandelbrot laws. The laws of Zipf and Mandelbrot are very important in information science. Indeed, they express very clearly how "skew" the distribution of the use of words in a text is. Although this is a property of all the so-called informetric laws (such as the ones of Lotka, Bradford, Leimkuhler and others - see e.g. Egghe and Rousseau (1990)), especially the laws of Zipf and Mandelbrot have important applications. Since these applications will also underline the importance of the laws of Zipf and Mandelbrot for multi-word phrases (to be studied in the sequel) we will briefly go into them.

It is well-known that, if symbols (e.g. letters, numbers) have an unequal chance of appearrance in texts, one can find unequal length coding such that the length becomes optimal (i.e. minimal) with respect to these unequal chances. In its purest form such a compression technique is called the Huffman compression technique, see Huffman (1952). The same can be said about words in texts. Their unequal appearrance in texts is ruled by the laws of Zipf or Mandelbrot. This knowledge leads to compression of texts; it also leads to the treatment of abbreviations. Such type of applications can be found in Heaps (1978), Aalbersberg, Brandsma and Corthout (1991), Schuegraf and Heaps (1973) and Jones (1979).

Knowing the exact form of the law of Zipf or Mandelbrot also enables researchers to draw conclusions on the stylistic type of texts. This in turn can lead to the proof (or disproval) of authorship or to the determination of the order in which different texts of the same author have been written. References to the first application are: Allen (1974), Brinegar (1963), McColly and Weier (1983), Michaelson and Morton (1971-1972), Smith (1983), Ule (1982) and Wickmann (1976). A reference to the second application is Cox and Brandwood (1959). A general treatment of these problems can be found in Herdan (1964).

The Laws of Zipf and Mandelbrot are also important in information retrieval (IR). Once known and used in the calculation of the entropy formula, one gets a measure of the selectivity of an IR result. This application to IR is intuitively clear: the law of Zipf-Mandelbrot deals with the unequal appearrance of words in texts and hence, indirectly by the "value" of key words in IR. In the same way, and since indexing and retrieval are similar ("dual") notions (see Egghe and Rousseau (1997)), the same can be said about the indexing value of key words. The law of Zipf-Mandelbrot is a basic ingredient in this in the sense that, via the calculation of the entropy, it determines the separation properties of key words. For this important application of automatic indexing see Salton and Mc Gill (1984).

The laws of Zipf and Mandelbrot are also important in the fractal aspects of information science and of linguistics. Fractals have been introduced by Mandelbrot (see Mandelbrot (1954, 1977a, b) in order to estimate the complexity of certain systems. The main idea behind fractals is the fact that, in real life, it is not possible to measure distances between objects since this is related with the scale of visualization of the object and since multiplying the scale by, say α, this leads to an increase of the distance larger than multiplying the old distance by α. A typical example is the problem of measuring the length of coast lines. For example (see Feder (1988)), put simply, if we multiply the scale of the map of the coast of Autralia by 2 we will find that the length has been multiplied by 1.52 times 2! D=1.52 is called the fractal dimension of the coastline and denotes, in a way, its complexity. Mandelbrot was able - under simplified circumstances - to calculate the fractal dimension D of a text and found that $D=\frac{1}{\beta}$, where β is the exponent appearing in his law (5). For a more understandable proof of this we refer the reader to Egghe and Rousseau (1990) - in fact its derivation is a simplified version of the argument that is given in section IV.

Let us finally mention applications of the laws of Zipf and Mandelbrot in speech recognition - see Chen (1989, 1991).

We hope to have indicated the incredible importance of the law of Zipf-Mandelbrot in the information sciences and in linguistics. In all of these applications one is limited to the consideration of single term key words. However multi-word phrases are very important objects in all of these applications and their importance is continuously increasing. It is e.g. clear that, from the point of view of compression of databases, multi-word phrases are more important than single words. Multi-word phrases are abbreviated more often than single words. The optimal allocation of abbreviations to such multi-word phrases, however, can only be studied if one knows the underlying law of Zipf-Mandelbrot for multi-word phrases.

Knowing these type of laws also opens perspectives in authorship determination: the use of multi-word phrases in texts reveals more of the style of an author than single terms; they have an "added value" above the value of knowing that the separated words have been used.

Perhaps the most important applications of laws of Zipf-Mandelbrot for multi-word phrases are to be expected in IR. This area experiences a fast evolution in IR-techniques, partially due to the steep increase of the importance of the Internet and with this of search engines (e.g. in WWW) to be used by, potentially, every person (the use of search engines is certainly not restricted to scientists or librarians as it was say two decades age when only "scientific" databases existed). With the steep increase of information one is in bad need of refined search techniques. One cannot suffice by using single words as key words. The separating possibility of multi-word keys are much higher than the cumulated one of the separate words. Knowing the law of Zipf-Mandelbrot for multi-word phrases might then lead to a scientific study of these "separation" values of multi-word phrases. As remarked above in the case of single key words, the law of Zipf-Mandelbrot for multi-word keys will lead us to a theory of automatic indexing of multi-word phrases.

Finally, knowing the type of law of Zipf-Mandelbrot will lead us to the determination of the complexity of texts, also considered to be composed of multi-word phrases, which gives a

better idea of the real nature of the text: a text is more than the set of its words; the study of multi-word phrases might help in the determination of the many complex aspects of texts.

Let us now investigate what type of distribuion is describing the occurrence of 2-word (3-word, ...) phrases in texts. As in the case of single words, all combinations of 2 (or more) consecutive words are taken into account. Smith and Devine (1985) stated that the occurrence of multi-word phrases can be described by distributions of the Zipf - Mandelbrot type (5). They furthermore found experimentally that the exponent β is decreasing the larger phrases one considers. So it was found that $\beta \approx 1$ for single words (i.e. classical Zipf - Mandelbrot), that $\beta \approx \frac{2}{3}$ for 2-word phrases and that $\beta \approx \frac{2}{5}$ for 3-word phrases. Similar observations were made by Hilberg (1989) and Meyer (1989 a,b). In these findings, English, German and even Chinese texts were the subject.

A rationale behind this was given by Mandelbrot for single words but was lacking for multi-word phrases until Rousseau (1997) studied the problem from a theoretical point of view. The argument is fractal, thereby extending Mandelbrot's original arguments (see Egghe and Rousseau (1990) from p.306 on for a detailed description of Mandelbrot's arguments).

To get an idea of what Rousseau is doing we will briefly introduce Mandelbrot's argument for single words. This will also be a good introduction to our derivation of the similar argument for 2-word phrases (given in section IV). Mandelbrot considers texts of words, formed by using an alphabet of N letters. He also assumes that each letter has an equal chance to occur, denoted by ρ. Words are formed by separating strings of letters by a blank. Blanks are possible by requiring that $\rho < \frac{1}{N}$. Hence $1-N\rho$ is the probability to have, at a certain spot in the text, a blank. By assuming independence between the letters in a word (which is a very much simplifying assumption) we then find that the probability for the occurrence of a word of length k (i.e. consisting of k letters) is

$$P_1(k) = \rho^k (1 - N\rho) \tag{6}$$

The same argument yields the probability to have a 2-word phrase consisting of k letters in total:

$$P_2(k) = \rho^k (1 - N\rho)^2. \qquad (7)$$

In total there are $(k-1)N^k$, 2-word phrases possible with k symbols and, although they will not all occur in the text, we do not follow the argument of Rousseau, when he assumes that only $\varepsilon_2 N^{k+s(2)}$ of them exist. The only reason why ε_2 is used (a constant as well as s(2)) is that he gets rid of the k-dependent factor k-1 in $(k-1)N^k$. There is no argument for such a "simplification" and in section IV we will in fact work with the formula $(k-1)N^k$ itself. Although indeed not all of these 2-word phrases occur, their probability of occurrence is governed by formula (7) and hence there is no need to add other (discrete) restrictions.

The main part of the paper, however, follows another approach. We will show that, supposing Mandelbrot's law to be valid for single words, that the corresponding law for multi-word phrases cannot be of this "power law" type. We will arrive at concrete expressions in the case of high ranks and supposing independance between the occurrence of the words. This is only a simplifying approximation in the same way as independence of occurrence of letters in single words was used in Mandelbrot and the same in multi-word phrases in Rousseau (1997) (admitting a correction factor in the latter paper).

So we arrive at the impossibility for the power law (7) to be valid in case of multi-word phrases and this is already found in the simplest models of independence between the words. We furthermore prove that if we approximate the found laws by a power law of type (7), that the $\beta_m s$ are decreasing with m. This represents a rationale for the statistical findings in Smith and Devine (1985), Hilberg (1989) and Meyer (1989 a,b).

As said, the paper closes with an examination of the fractal argument of Mandelbrot in the case of 2-word phrases. Here we do not use the assumptions of Rousseau. We show that the same type of law is found as in the above described approach, giving another rationale for it.

II. The case of 2-word phrases.

Although we have a general solution for m-word phrases (m=2,3...) we will deal with the 2-word case separately. The reason is that the general argument is rather intricate (see appendix A) but that this argument is similar to the much simpler one in the 2-word case.
Denote by

$$P_1(r) = \frac{C}{r^{\beta}} \tag{8}$$

Zipf - Mandelbrot's law, which is supposed to be valid in the single word case. Hence $P_1(r)$ denotes the probability of occurrence of a word that has rank r (ranking according to single words, of course). We use independence of occurrence of two consecutive words as a simplification (we note again that also in Mandelbrot, in order to obtain (8), independence between letters is used as a simplification). We hence have

$$P_2(r_1, r_2) = P_1(r_1)P_1(r_2), \tag{9}$$

where $P_2(r_1,r_2)$ denotes the probability of occurrence of a 2-word phrase that consists of a word with rank r_1 and one with rank r_2 in the single word case.

Note that $\sum_{r_1,r_2} P_2(r_1,r_2) = 1$ since $\sum_{r} P_1(r) = 1$.

We have solved our problem if we can find an estimate of the final rank r of this 2-word phrase, i.e. find r such that

$$P_2(r) = P_2(r_1, r_2), \tag{10}$$

where r denotes the ranking of a 2-word phrase consisting of a word with rank r_1 and one with rank r_2 in the single word case (8). It is already certain that this is not an ill-posed problem: r_1 and r_2 determine $P_1(r_1)$ and $P_1(r_2)$, hence (9), hence the ranking according to the value of

$P_2(r_1,r_2)$, hence r (although ties are possible). We will only be able to solve this problem in case r_1 and r_2 are high. In any case we have

$$r = \#\{(r'_1, r'_2) \parallel P_1(r'_1)P_1(r'_2) \geq P_1(r_1)P_1(r_2)\} \tag{11}$$

(# = number of elements in).

The inequality

$$P_1(r'_1)P_1(r'_2) \geq P_1(r_1)P_1(r_2) \tag{12}$$

is equivalent with, using (8),

$$r_1 r_2 \geq r'_1 r'_2 \tag{13}$$

(Note that $B=\frac{N-1}{N}$ as follows from the fractal argument of Mandelbrot, but we do not need this result here). We put $r'_1=r_1+i$, count the number of possible i's and then the value of

$$r'_2 \leq \frac{r_1 r_2}{r_1 + i} \quad , \tag{14}$$

according to (13). Only i-values that yield

$$\frac{r_1 r_2}{r_1 + i} \geq 1$$

are possible since r'_2 is a rank. Hence

$$i \leq r_1(r_2 - 1) \tag{15}$$

Of course, also r'_1 is a rank, hence

$$i \geq -(r_1 - 1) \tag{16}$$

(15), (16) and (14) yield in (11)

$$r \approx \sum_{i=-(r_1-1)}^{r_1(r_2-1)} \frac{r_1 r_2}{r_1 + i} \tag{17}$$

But, using the integral test for series,

$$\sum_{i=-(r_1-1)}^{r_1(r_2-1)} \frac{1}{r_1 + i}$$

$$\approx \int_{-(r_1-1)}^{r_1(r_2-1)} \frac{di}{r_1 + i} = \ell n(r_1 r_2) \quad . \tag{18}$$

(17) and (18) yield

$$r \approx r_1 r_2 \ell n(r_1 r_2) \quad . \tag{19}$$

Formulae (8), (9), (10) and (19) now yield

$$P_2(r) \approx P_2\left(r_1 r_2 \ell n(r_1 r_2)\right) \approx \frac{C^2}{(r_1 r_2)^{\beta}} \tag{20}$$

Denote by $\varphi(x)$ the inverse function of the injective function $x \to x\ln x$,then we have that

$$P_2(r) = \frac{C^2}{(\varphi(r))^{\beta}} \tag{21}$$

clearly indicating that, even asymptotically, P_2 - i.e. the analogue of Zipf - Mandelbrot's law for 2-word phrases - cannot be a power function as in the case (4) or (5).

However, forcing (21) into a power law (in r), as is the case with statistical fitting of data, shows that the exponent β' in this new law is inferior to β . Indeed, if we write

$$\frac{C^2}{(\varphi(r))^{\beta}} \approx \frac{D}{r^{\beta'}} \tag{22}$$

where also D is a constant, yields

$$\beta' \approx \frac{\ell n\left(\frac{D}{C^2}\right)}{\ell n r} + \beta \frac{\ell n \varphi(r)}{\ell n r} \tag{23}$$

Hence, in terms of r_1 and r_2 and by definition of φ we have

$$\beta' \approx \frac{\ell n\left(\frac{D}{C^2}\right)}{\ell n\left(r_1 r_2 \ell n(r_1 r_2)\right)} + \beta \frac{\ell n(r_1 r_2)}{\ell n\left(r_1 r_2 \ell n(r_1 r_2)\right)}$$

So

$$\beta' \approx \frac{\ell n\left(\frac{D}{C^2}\right)}{\ell n\left(r_1 r_2 \ell n(r_1 r_2)\right)} + \beta \frac{1}{1 + \frac{\ell n\left(\ell n(r_1 r_2)\right)}{\ell n(r_1 r_2)}} \tag{24}$$

Since r_1 and r_2 are large, the first term can be made as small as we wish. Furthermore is, in the second term, the value of the coefficient of β clearly below 1. In conclusion: $\beta' < \beta$ in fitting. Note that we cannot put an equality in formula (22) since this would then not result in a power law $\frac{D}{r^{\beta'}}$ with D a constant. We can from (24) also conclude that, since r_1 and r_2 are high, that

$$\beta' \approx \beta \frac{1}{1 + \frac{\ell n\left(\ell n(r_1 r_2)\right)}{\ell n(r_1 r_2)}} \; . \tag{25}$$

Summarizing this section on 2-word phrases we have the following results:

Theorem:

Under the assumption of independence of the occurrence of consecutive words we obtain the following law for the occurrence of 2-word phrases in texts. Let $P_2(r)$ denote the probability that a two-word phrase on rank r occurs. Then, if r is large,

$$P_2(r) = \frac{C^2}{(\varphi(r))^{\beta}}$$

where φ is the inverse of the function $x \to x\ln x$ and where we suppose the law of Zipf - Mandelbrot

$$P_1(r_1) = \frac{C}{r_1^{\beta}}$$

to be valid for single words. The rank r for 2-word phrases is obtained from the ranks r_1 and r_2 of the single words via the formula (for large r_1 and r_2)

$$r \approx r_1 r_2 \ell n(r_1 r_2)$$

(hence $\varphi(r) \approx r_1 r_2$).

Finally, when approximating (21) by a power of the type (22) we find that

$$\beta' \approx \beta - \frac{1}{1 + \frac{\ell n(\ell n(r_1 r_2))}{\ell n(r_1 r_2)}} < \beta \quad .$$

III. The case of m-word phrases.

In this section we study the probability of occurrence of m-word phrases (m=2,3,4,...). The general case is similar to the case m=2 though a lot more intricate. We again suppose the law of Zipf - Mandelbrot (8) to be valid for single words and we use high ranks.
In appendix A we prove the following result

Theorem:
Let $P_m(r)$ denote the probability that an m-word phrase on rank r occurs. Then, if r is large,

$$P_m(r) = \frac{C^m}{(\varphi_m(r))^{\beta}} \tag{26}$$

where φ_m is the inverse of the function

$$x \to \frac{x\ell n^{m-1}x}{m-1} \tag{27}$$

This rank r for m-word phrases is obtained from the ranks r_1, r_2,...,r_m of the single words via the formula (for large r_1,...,r_m)

$$r \approx \frac{r_1 \cdots r_m}{m-1} \ell n^{m-1}(r_1 \ldots r_m) \tag{28}$$

(hence $\varphi_m^{(r)} = r_1 \ldots r_m$). Finally, when appoximating (26) by a power law of type

$$\frac{D}{r^{\beta' m}} \tag{29}$$

we have that

$$\beta'_m \approx \beta \frac{1}{1 + \frac{\ell n\left(\frac{\ell n^{m-1}(r_1 \ldots r_m)}{m-1}\right)}{\ell n(r_1 \ldots r_m)}} < \beta \quad . \tag{30}$$

If we take $r_1 \approx \ldots \approx r_m$ in the above formula, it can be shown that the sequence $\beta'_2, \beta'_3, \ldots$ is strictly decreasing.

All these findings are in accordance with the experimental findings in Smith and Devine (1985), Hilberg (1989) and Meyer (1989 a,b), but we stress the fact that the Zipf - Mandelbrot exponents β'_m only apply to a power law that approximates the non-power law (26). So we have presented a rationale for these experimental findings and found that the power law of Zipf - Mandelbrot is not correct for the case of m-word phrases. The formula

$$r \approx \frac{r_1 \ldots r_m}{m-1} \ell n^{m-1}(r_1 \ldots r_m) \tag{28}$$

has independent interest: it gives a formula for the rank r of an m-word phrase in function of the ranks $r_1 \ldots r_m$ of the single words, supposing the Zipf - Mandelbrot law for single words. Note that this formula is independent of the parameters C and β in this law and only uses the fact that a power law applies for single words! This finding in itself makes it clear that a classical Zipf - Mandelbrot power law cannot apply in the case of m-word phrases (m=2,3,...).

IV. The fractal argument of Mandelbrot for 2-word phrases.

We suppose that we have an alphabet of N letters. Words are formed with these letters and these words are separated by a blank. In total there are $(k\text{-}1)N^k$ possible 2-word phrases consisting of k letters in total. We put ρ as the probability for a letter to occur. Since there are also blanks we have that $\rho < \frac{1}{N}$. The probalility to have a 2-word phrase consisting of k letters hence is

$$P_2 = \rho^k (1 - N\rho)^2 \tag{29}$$

(1-Nρ is the probability to have a blank).

Since this is decreasing in k we can estimate the rank r of this word as follows (cfr. the Mandelbrot argument for single words - see e.g. Egghe and Rousseau (1990))

$$\begin{aligned} & N^2 + 2N^3 + \ldots + (k-2)N^{k-1} \\ & = \frac{N}{(N-1)^2}(iN^{i+1} - (i+1)N^i + 1) \end{aligned} \tag{30}$$

Indeed, 2-word phrases consist of at least 2 letters[2]. A long calculation but only involving sums of geometric series yields

$$\begin{aligned} & N + 2N^2 + \ldots + iN^i \\ & = \frac{N}{(N-1)^2}\left(iN^{i+1} - (i+1)N^i + 1\right) \end{aligned} \tag{31}$$

for all i. Applied in (30) this yields

$$\begin{aligned} & (k-2)N^{k-1} - (k-1)N^{k-2} < \left(\frac{N-1}{N}\right)^2 r - 1 \\ & \leq (k-1)N^k - kN^{k-1} \end{aligned} \tag{32}$$

To fix r we will take the average of both sides:

$$\begin{aligned} \left(\frac{N-1}{N}\right)^2 r - 1 &\approx \frac{1}{2}\left[(k-2)N^{k-1} - (k-1)N^{k-2} + (k-1)N^k - kN^{k-1}\right] \\ &= \frac{1}{2}N^{k-2}\left[\left(N^2-1\right)(k-1) - 2N\right] \\ \left(\frac{N-1}{N}\right)^2 r - 1 &\approx \frac{1}{2}N^{k-2}\left(N^2-1\right)(k-1) \end{aligned} \tag{33}$$

This is the (average) rank of 2-word phrases consisting of k letters in total. From (29) it follows that

$$P_2(r) = \rho^k(1-N\rho)^2 = \rho^{k-1}\rho(1-N\rho)^2$$

---

[2] This also corrects the argument in Egghe and Rousseau (1990)), p. 306 for single words : one should start with N instead of 1 in the analogous inequality since words have at least 1 letter. This correction yields $\frac{N-1}{N}$ instead of N as the coefficient of r in the formula for P.

Hence

$$k-1=\frac{\ell n\left(\dfrac{P_2(r)}{\rho(1-N\rho)^2}\right)}{\ell n\rho} \quad . \tag{34}$$

(34) in (33) yields

$$\ell n\left[\left(\frac{P_2(r)}{\rho(1-N\rho)^2}\right)^{\frac{\ell nN}{\ell n\rho}}\right].\left(\frac{P_2(r)}{\rho(1-N\rho)^2}\right)^{\frac{\ell nN}{\ell n\rho}}$$

$$=\frac{2N\ell nN}{N^2-1}\left[\left(\frac{N-1}{N}\right)^2 r-1\right]$$

$$\approx\frac{2(\ell nN)(N-1)}{N(N+1)}r \tag{35}$$

for large r. So

$$\left(\frac{P_2(r)}{\rho(1-N\rho)^2}\right)^{\frac{\ell nN}{\ell n\rho}}=\varphi(\alpha r) \tag{36}$$

with

$$\alpha=\frac{2(\ell nN)(N-1)}{N(N+1)} \quad ,$$

a fixed number and where φ is (as in formula (21)) the inverse of the function $x\to x\ell nx$.

We finally have

$$P_2(r) = \frac{D}{(\varphi(\alpha r))^{\beta}} \tag{37}$$

where

$$D = \rho(1 - N\rho)^2$$

and

$$\beta = -\frac{\ell n \rho}{\ell n N} \tag{38}$$

(cfr. $\beta = \frac{1}{D_s}$, the fractal dimension of the text, cf. Egghe and Rousseau (1990), p.307).

Up to the appearance of α in (36), this formula is exactly the same as the one obtained in section II (formula (21)). For high r we can even get rid of α as follows: if we put

$$x = \left(\frac{P_2(r)}{\rho(1 - N\rho)^2} \right)^{\frac{\ell n N}{\ell n \rho}} \tag{39}$$

then (35) reads

$$x \ell n x = \alpha r \tag{40}$$

So

$$r = \frac{1}{\alpha} x \ell n x$$

$$= \frac{1}{\alpha} x \ell n \left(\frac{1}{\alpha} x \right) - \frac{1}{\alpha} x \ell n \frac{1}{\alpha}$$

$$r \approx \frac{1}{\alpha} x \ell n \left(\frac{1}{\alpha} x \right) \quad , \tag{41}$$

hereby using that

$$\frac{1}{\alpha} x\ell n \frac{1}{\alpha} << \frac{1}{\alpha} x\ell n\left(\frac{1}{\alpha} x\right)$$

since

$$\frac{1}{\alpha} x\ell n\left(\frac{1}{\alpha} x\right)$$

$$= \frac{1}{\alpha} x\ell n \frac{1}{\alpha} + \frac{1}{\alpha} x\ell n x >> \frac{1}{\alpha} x\ell n \frac{1}{\alpha} \quad .$$

Here we use that r, hence by (40) x, is large (α is fixed). From (41) it follows that

$$\frac{1}{\alpha} x \approx \varphi(r) \tag{42}$$

Hence

$$\frac{1}{\alpha}\left(\frac{P_2(r)}{\rho(1-N\rho)^2}\right)^{\frac{\ell nN}{\ell n\rho}} \approx \varphi(r) \ ,$$

yielding

$$P_2(r) \approx \frac{E}{(\varphi(r))^{\beta}} \tag{43}$$

with β still as in (38) and with

$$E = \rho(1-N\rho)^2 \alpha^{\frac{\ell n\rho}{\ell nN}} \quad . \tag{44}$$

Formula (43) is exactly formula (21).

This gives a second rationale for the validity of (21) or (43) in the case of 2-word phrases, hence disproving again the validity of the Zipf - Mandelbrot power law.

Analogus arguments can be given for m-word phrases. We leave this to the reader.

APPENDIX A

Proof of the theorem in section III on the probability of occurrence and on the ranks of *m*-word phrases.

Let the *m*-word phrase consist of *m* single words with ranks (amongst the single words) of occurrence $r_1,\ldots r_m$. Supposing independence as before, we obtain the rank r of the *m*-word (amongst all the *m*-words) as the number of vectors $(r'_1,\ldots,r'_m)$ for which

$$P_1(r'_1)\ldots P_1(r'_m) \geq P_1(r_1)\ldots P_1(r_m), \tag{A1}$$

where

$$P_1(x) = \frac{C}{x^{\beta}} \tag{A2}$$

is the classical Zipf - Mandelbrot law for single words. (A1) and (A2) yield the condition

$$r'_1 \ldots r'_m \leq r_1 \ldots r_m \tag{A3}$$

We have to count all possibilities. For r'_m this is

$$r'_m \leq \frac{r_1 \ldots r_m}{r'_1 \ldots r'_{m-1}} \tag{A4}$$

and for $r'_1,\ldots,r'_{m-1}$ we put

$$r_\ell' = r_\ell + k_\ell \tag{A5}$$

$l=1,\ldots,m-1$. Of course

$$k_\ell \geq -(r_\ell - 1) \tag{A6}$$

for all $l=1,\ldots,m-1$ since ranks must be larger than or equal to 1. For the same reason (A4) implies

$$\frac{r_1 \dots r_m}{(r_1 + k_1) \dots (r_{m-1} + k_{m-1})} \geq 1$$

hence

$$\frac{r_1 \dots r_m - (r_1 + k_1) \dots (r_{m-2} + k_{m-2}) r_{m-1}}{(r_1 + k_1) \dots (r_{m-2} + k_{m-2})} \geq k_{m-1} \tag{A7}$$

Applying (A7) and (A6) for l=m-1 yields

$$\frac{r_1 \dots r_m - (r_1 + k_1) \dots (r_{m-3} + k_{m-3}) r_{m-2}}{(r_1 + k_1) \dots (r_{m-3} + k_{m-3})} \geq k_{m-2} \tag{A8}$$

Continuing in this way we obtain

$$k_\ell \leq \frac{r_1 \dots r_m - (r_1 + k_1) \dots (r_{\ell-1} + k_{\ell-1}) r_\ell}{(r_1 + k_1) \dots (r_{\ell-1} + k_{\ell-1})} \tag{A9}$$

for all l=2,...m-1. Finally, using this for l=2 and (A6) for l=2 we obtain

$$k_1 \leq r_1 (r_2 \dots r_m - 1) \tag{A10}$$

The formulae (A5), (A9) and (A10) give us all possibilities and hence, using (A4) we obtain

$$r \approx r_1 \dots r_m \sum_{k_1} \frac{1}{r_1 + k_1} \dots \sum_{k_{m-1}} \frac{1}{r_{m-1} + k_{m-1}} \tag{A11}$$

where each k_l ranges between the values indicated in (A6) and (A9) and (A10) (for l=1). Denoting the upper values in (A9) by α_l and approximating the $\sum$ by integrals (using the integral test for series), we find

$$r \approx r_1 \dots r_m \int_{-(r_1 - 1)}^{r_1 (r_2 \dots r_m - 1)} \frac{dk_1}{r_1 + k_1} \dots \int_{-(r_{m-\ell'} - 1)}^{\alpha_{m-\ell'}} \frac{dk_{m-\ell'}}{r_{m-\ell'} + k_{m-\ell'}}$$

$$\dots \int_{-(r_{m-1} - 1)}^{\alpha_{m-1}} \frac{dk_{m-1}}{r_{m-1} + k_{m-1}} \tag{A12}$$

where l' denotes an arbitrary value between 1 and m-1 (these cases are given explicitely). This intricate form is calculated as follows.

A. Case $l'=1$

$$\int_{-(r_{m-1}-1)}^{\alpha_{m-1}} \frac{dk_{m-1}}{r_{m-1}+k_{m-1}}$$

$$= \ell n\left(r_{m-1}+k_{m-1}\right) \Big]_{k_{m-1}=-(r_{m-1}-1)}^{k_{m-1}=\alpha_{m-1}}$$

with α_{m-1} given by the upper bound in (A9) for $l=m$-1. This gives the value

$$\ell n \frac{r_1 \ldots r_m}{\left(r_1+k_1\right)\ldots\left(r_{m-2}+k_{m-2}\right)}.$$

After calculating this way the further factors it becomes clear what the general formula is going to be. We will prove it by complete induction: suppose that we have the result

$$\frac{1}{\ell'-1}\ell n^{\ell'-1}\left(\frac{r_1 \ldots r_m}{\left(r_1+k_1\right)\ldots\left(r_{m-\ell'}+k_{m-\ell'}\right)}\right)$$

after l'-1 steps. Step l' is then

$$\int_{-(r_{m-\ell'}-1)}^{\alpha_{m-\ell'}} \frac{\frac{1}{\ell'-1}\ell n^{\ell'-1}\left(\frac{r_1 \ldots r_m}{\left(r_1+k_1\right)\ldots\left(r_{m-\ell'}+k_{m-\ell'}\right)}\right)}{r_{m-\ell'}+k_{m-\ell'}} dk_{m-\ell'}$$

$$= \int_{-(r_{m-\ell'}-1)}^{\alpha_{m-\ell'}} \frac{\frac{1}{\ell'-1}\ell n^{\ell'-1}\left(\frac{r_1 \ldots r_m}{\left(r_1+k_1\right)\ldots\left(r_{m-\ell'-1}+k_{m-\ell'-1}\right)}\right)}{r_{m-\ell'}+k_{m-\ell'}} dk_{m-\ell'}$$

$$- \int_{-(r_{m-\ell'}-1)}^{\alpha_{m-\ell'}} \frac{1}{\ell'-1} \frac{\ell n^{\ell'-1}\left(r_{m-\ell'}+k_{m-\ell'}\right)}{r_{m-\ell'}+k_{m-\ell'}} dk_{m-\ell'}$$

where $\alpha_{m-l'}$ is given as the upper bound in (A9) for $l=m-l'$ .This gives the result

$$\frac{1}{\ell'}\ell n^{\ell'}\left(\frac{r_1\ldots r_m}{(r_1+k_1)\ldots(r_{m-\ell'-1}+k_{m-\ell'-1})}\right)$$

concluding the induction step. By (A12) and the above result we can now conclude that

$$r \approx \frac{r_1\ldots r_m}{m-1}\ell n^{m-1}(r_1\ldots r_m) \quad . \tag{A13}$$

Since we have, by definition of $P_m(r)$, that

$$P_m(r) = P_1(r_1)\ldots P_1(r_m) \tag{A14}$$

we have now that

$$P_m\left(\frac{r_1\ldots r_m}{m-1}\ell n^{m-1}(r_1\ldots r_m)\right)$$

$$= \frac{C^m}{(r_1\ldots r_m)^\beta} \tag{A15}$$

or else, by (A13)

$$P_m(r) = \frac{C^m}{(\varphi_m(r))^\beta} \tag{A16}$$

where φ_m is the inverse function of the injective function

$$x \to \frac{x\ell n^{m-1}(x)}{m-1} \tag{A17}$$

An analogous argument as in the case m=2 now yields that, if we put

$$\frac{C^m}{(\varphi_m(r))^\beta} \approx \frac{D_m}{r^{\beta' m}} \tag{A18}$$

(cfr.(22)), we find

$$\beta'_m \approx \beta \frac{1}{1+\frac{\ell n\left(\frac{\ell n^{m-1}(r_1 \ldots r_m)}{m-1}\right)}{\ell n(r_1 \ldots r_m)}} \quad . \tag{A19}$$

It is clear that $\beta'_m<\beta$ for every m = 2,3,.... Since formula (A19) is valid for all $r_1,\ldots,r_m$ which are large, we can study the behavior of (A19) in case $r_1 \approx \ldots \approx r_m$ (denoted as r_0) and in case r_0 is large. It is clear that the sequence β'_m decreases if the sequence α_m increases, where

$$\alpha_m = \frac{\ell n\left(\frac{\ell n^{m-1}(r_1 \ldots r_m)}{m-1}\right)}{\ell n(r_1 \ldots r_m)} \tag{A20}$$

Now we have

$$\alpha_m = \frac{\ell n\left(\frac{\ell n^{m-1}\left(r_0^m\right)}{m-1}\right)}{\ell n\left(r_0^m\right)}$$

$$\alpha_m = \frac{(m-1)\ell n(m \ell n r_0) - \ell n(m-1)}{m \ell n r_0} \tag{A21}$$

Now $\alpha_{m+1}>\alpha_m$ iff

$$\ell n r_0 > \frac{m^{m^2+m-1}}{(m-1)^{m+1}(m+1)^{m^2}} \tag{A22}$$

which is satisfied for all $r_0 \geq 3$ on (and this is so since we supposed r_0 to be large). Indeed the critical r_0 - values for $\alpha_{m+1}>\alpha_m$ are (m=2,3,4,5,...)

$\alpha_{m+1} > \alpha_m$	from r_0 on
$m = 2$	2.480
$m = 3$	1.100
$m = 4$	1.020
$m = 5$	1.004
⋮	⋮

This concludes the proof that the sequence β'_m is decreasing in m.

References

Aalbersberg, IJ., Brandsma, E. and Corthout, M. (1991). Full-text document retrieval: from theory to applications. In: Informatiewetenschap 1991 (Kempen and de Vroomen, eds.); Leiden: STINFON, 3-17.

Allen, J.R. (1974). Methods of author identification through stylistic analysis. The French Review, XLVII (5), 904-916.

Brinegar, C. (1963). Mark Twain and the Quintus Curtius Snodgrass letters: a statistical test of authorship. Journal of the American Statistical Association, 58, 85-96.

Chen, Y.-S. (1989). Zipf's low in text modeling. International Journal of General Systems, 15, 233-252.

Chen, Y.-S. (1991). Statistical models of text in continuous speech recognition. Kybernetes, 20 (5), 29-40.

Cox, D.R. and Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. Journal of the Royal Statistical Society, B, 21, 195-200.

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

Egghe, L. and Rousseau, R. (1997). Duality in information retrieval and the hypergeometric distribution. Journal of Documentation, 53 (5), 488-496.

Feder, J. (1988). Fractals. Plenum, New York.

Heaps, H.S. (1978). Information Retrieval: computational and theoretical aspects. Academic Press, New-York.

Herdan, G. (1964). Quantitative linguistics. Butterworths, London.

Hilberg, W. (1988). Das Netzwerk der menschlichen Sprache und Grundzüge einer entsprechend gebauten Sprachmaschine. ntzArchiv, 10, 133-146.

Huffman, D.A. (1952). A method for the construction of minimum redundancy codes. Prodeedings of the IRE, 1098-1101.

Jones, D.S. (1979). Elementary Information Theory. Oxford Applied Mathematics and Computing Series, Clarendon Press, Oxford.

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

Mandelbrot, B.B. (1977a). Fractals, Form, Chance and Dimension. Freeman, San Fransisco.

Mandelbrot, B.B. (1977b). The Fractal Geometry of Nature. Freeman, New York.

Mc Colly, W. and Weier, D. (1983). Literary attribution and likelihood-ratio tests: the case of the middle English PEARL-poems. Computers and Humanities, 17, 65-75.

Meyer, J. (1989a). Gilt das Zipfsche Gesetz auch für die chinesische Schriftsprache? ntzArchiv, 11, 13-16.

Meyer, J. (1989 b). Statistische Zusammenhänge in Texten. ntzArchiv, 11, 287-294.

Michaelson, S. and Morton, A.Q. (1971-1972). Last words. A test of authorship for Greek authors. New Testament Studies, 18, 192-208.

Rousseau, R. (1997). A fractal approach to word occurrences in texts: the Zipf - Mandelbrot law for multi-word phrases. Preprint.

Salton, G. and Mc Gill, M.J. (1984). Introduction to modern Information Retrieval. Mc Graw-Hill, Auckland.

Schuegraf, E.J. and Heaps, H.S. (1973). Selection of equifrequent word fragments for information retrieval. Information Storage and Retrieval, 9, 697-711.

Smith, F.J. and Devine, K. (1985). Storing and retrieving word phrases. Information Processing and Management, 21, 215-224.

Smith, M.W.A. (1983). Recent experience and new developments of methods for the determination of authorship. Association for Literary and Linguistical Computing Bulletin, 11 (3), 73-82.

Ule, L. (1982). Recent progress in computer methods of authorship determination. Association for Literary and Linguistical Computing Bulletin, 10 (3), 73-89.

Wickmann, D. (1976). On disputed authorship, statistically. Association for Literary and Linguistical Computing Bulletin 4 (1), 32-41.

Zipf, G. K. (1949). Human Behavior and the Principle of least Effort. Addison - Wesley, Cambridge. Reprinted in 1965, Hafner, New-York.