The exact rank-frequency function and size-frequency function of N-grams and N-word phrases with applications

Peer-reviewed author version

EGGHE, Leo (2005) The exact rank-frequency function and size-frequency function of N-grams and N-word phrases with applications. In: MATHEMATICAL AND COMPUTER MODELLING, 41. p. 807-823.

DOI: 10.1016/j.mcm.2003.12.016
Handle: http://hdl.handle.net/1942/743

The exact rank-frequency function and size-frequency function of N-grams and N-word phrases with applications

by

L. Egghe

Limburgs Universitair Centrum (LUC), Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@luc.ac.be

ABSTRACT

N-grams are generalized words consisting of N consecutive symbols (letters), as they are used in a text. N-word phrases are general concepts consisting of N consecutive words, also as used in a text. Given the rank-frequency function of single letters (i.e. 1-grams) or of single words (i.e. 1-word phrases) being Zipfian, we determine in this paper the exact rank-frequency function (i.e. the occurrence of N-grams or N-word phrases on each rank) and size-frequency distribution (i.e. the density of N-grams or N-word phrases on each occurrence density) of these N-grams and N-word phrases. This paper distinguishes itself from other ones on this

[1] Permanent address: Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium.

Key words and phrases: N-gram, N-word phrase, rank-frequency distribution, size-frequency distribution, Zipfian distribution.

topic by allowing no approximations in the calculations. This leads to an intricate rank-frequency function for N-grams and N-word phrases (as we knew before from unpublished calculations) but leads surprisingly, to a very simple size-frequency function f_N for N-grams or N-word phrases of the form

$$f_N(j) = \frac{F}{j^{1+\frac{1}{\beta}}} \ln^{N-1}\left(\frac{G}{j}\right)$$

where the Zipfian distribution of single letters or words is proportional to $\frac{1}{r^\beta}$.

The paper closes with the calculation of type/token averages μ_N and type/token-taken averages μ_N^* for N-grams and N-word phrases, where we also verify the theoretically proved result $\mu_N^* \geq \mu_N$ but where we also give estimates for the differences $\mu_N^* - \mu_N$.

I. Introduction

N-grams and N-word phrases are very important objects in information science. This is obvious for N-word phrases, being the basis for linguistical expression and allow for more complex ideas than the single words on their own. Because of this importance, N-word phrases are indexed as separate entities (pre-coordinative indexing) and this implies that their use in information retrieval (IR) (post-coordinative retrieval) is basic in the refinement of searches.

N-grams, as indicated in Egghe (2000a) have important applications in indexing and IR (generalizing e.g. truncation, useful in any language but especially in Asian languages where, because of their special structure, truncation is not so efficient), error detection and correction, text compression, identification of languages or of authorship, subject classification and even speech recognition and the indexing and retrieval of music. For more details on these applications we refer the reader to Cohen (1995), Damashek (1995), Robertson and Willett

(1998), Grossman and Frieder (1998) (Section 3.4), Yannakoudakis, Tsomokos and Hutton (1990) and Nelson and Downie (2001).

Because of the importance of N-grams and N-word phrases, their informetric properties should be revealed. It is clear that these N-tuples can be considered as elements of an N-fold Cartesian product of the space of the single objects, being respectively single letters and single words. These single objects have well-established informetric properties: basically their rank-frequency distributions can be described by the law of Zipf (see Zipf (1949), Herdan (1960), Egghe and Rousseau (1990) or Baayen (2001)), which is a power law of the form

$$P_1(r) = \frac{B}{r^{\beta}} \tag{1}$$

where $r^{3}\ 1$. This is well-known in linguistics (for single words) and shown to be applicable to the distribution of single letters in Egghe (2000a).

A first attempt to derive the rank-frequency function for N-word phrases was given in Egghe (1999) but using a lot of simplifying assumptions and approximations. A substantial improvement has been given in Egghe (2000a) where the argument was also applicable to general N-grams. The calculation of the general rank-frequency function is very tedious and for this reason, in Egghe (1999) as well as in Egghe (2000a), a technical simplification (approximation) has been adopted on the rank ranges of the single letters or words (we will indicate exactly what type of simplification that was used).

In this paper we drop this simplification, leading to an intricate rank-frequency function for N-grams and N-word phrases. Surprisingly, however, when calculating the size-frequency function which is equivalent with the obtained rank-frequency function for N-grams and N-word phrases, we obtain a very simple expression (even much simpler than the one obtained from the simplified rank-frequency argument!): supposing (1) to be valid for single letters or words (i.e. $N = 1$) we will show in this paper that the size-frequency function f_N of N-grams or N-word phrases has the form

$$f_N(j)=\frac{F}{j^{1+\frac{1}{\beta}}}\ln^{N-1}\left(\frac{G}{j}\right) \qquad (2)$$

Note that, for $N = 1$, (2) conforms with the known law of Lotka

$$f_1(j)=\frac{F}{j^{1+\frac{1}{\beta}}}=\frac{F}{j^{\alpha}} \qquad (3)$$

where $\alpha = 1+\frac{1}{\beta}$ is Lotka's exponent (see Egghe (1989, 1990) or Egghe and Rousseau (1990)) and where f_1 is the law of Lotka, known to be equivalent with Zipf's law (cf. Rousseau (1990)).

The simple form (2) then enables us to derive formulae for the average number of occurrences of N-grams and N-word phrases and for the average number of uses of these N-tuples (Type/Token-Taken informetrics as described in Egghe (2003)).

In the next section we will repeat the basic formulae of informetrics on rank and size-frequency functions and their interrelations (Type/Token informetrics) and we will also repeat the basic facts of Type/Token-Taken informetrics.

Section III is then devoted to the intricate correct calculation of the rank-frequency function of N-grams and N-word phrases, using Zipf's law (1) for the $N = 1$ case.

Section IV then derives from this the simple size-frequency function (2) and Section V applies the latter result to the calculation of formulae of average occurrence μ_N and average use μ_N^* of these N-tuples. These formulae are also calculated in practise and the results compared.

II. Basic formulae in classical informetrics.

We refer the reader to Egghe (1989, 1990), Egghe and Rousseau (1990), Rousseau (1990), for more details on the following definitions and results. Basic in informetrics theory is an informetric production process (IPP) in which one has sources producing (or having) items (e.g. authors or journals produce papers, papers produce references or citations,…). The basic informetric function is the function $f : j \rightarrow f(j)$ where $j \in [1, \rho_m]$ and where $f(j)$ denotes the density of sources with item density j and where ρ_m denotes the maximal item density. It is the continuous size-frequency function (of which the Lotka power law is an example). The rank-frequency function $g : r \rightarrow g(r)$, where $r \in [0, T]$ expresses the item-density in the source on rank-density r and where T denotes the total number of sources.

The functions f and g relate as follows (g^{-1} denotes the inverse of g):

$$g^{-1}(j) = r(j) = \int_j^{\rho_m} f(j')dj' \tag{4}$$

and also

$$f(j) = -\frac{1}{g'\left(g^{-1}(j)\right)}. \tag{5}$$

We also have that

$$\int_0^T g(r)dr = A, \tag{6}$$

the total number of items, hence

$$P(r) = \frac{g(r)}{A} \tag{7}$$

is the actual rank-frequency distribution.

In this framework, the law of Zipf is given by (since $r \in [0,T]$)

$$g(r) = \frac{E}{(1+r)^{\beta}} \tag{8}$$

which boils down to (1), replacing r by $r' = 1 + r \in [1, T+1]$. The distribution-form of the above Zipf function is given by, using (7):

$$P(r) = \frac{D}{r^{\beta}} \tag{9}$$

where $D = \frac{E}{A}$ and where $r \in [1, T+1]$ (we, henceforth, drop the primes in r').

It follows from (4) that

$$T = \int_{1}^{\rho_m} f(j)dj \tag{10}$$

and it can easily been proved that

$$A = \int_{1}^{\rho_m} jf(j)dj\,. \tag{11}$$

Hence

$$\mu = \frac{A}{T} = \frac{\int_{1}^{\rho_m} jf(j)dj}{\int_{1}^{\rho_m} f(j)dj} \tag{12}$$

denotes the average number of items per source (i.e. as they exist or occur). This is also called the Type/Token (TT) average (using terminology from linguistics). In Egghe (2003), Type/Token-Taken (TTT) informetrics is developed where also the use of the items is taken into account. Let us just give one example (further examples and applications can be found in Egghe (2003)): N-grams of books occur in a database (e.g. an OPAC) and describing this occurrence (incl. the average number μ that an N-gram occurs in this database) is the domain of TT-informetrics. A cataloguer (e.g.), using this database to check whether or not a new book (that has to be catalogued) is already in the catalogue will enter the corresponding N-gram of this book. Hence, the more an N-gram occurs in the catalogue, the more it will also be typed in the retrieval process (assuming that N-grams of already catalogued books have the same distribution as N-grams of books that have to be catalogued). The informetrics describing this "use" of items is called TTT-informetrics and it is proved in Egghe (2003) that its size-frequency function, denoted f^*, is given by

$$f^*(j) = jf(j) \qquad (13)$$

for all $j \in [1, \rho_m]$ (note that the item densities – in practise number of times an N-gram occurs in the catalogue) remain the same in TT- as in TTT-informetrics. Based on (10), (11) and (12), we have that the TTT-average, denoted μ^*, is now given by

$$\mu^* = \frac{A}{W} \qquad (14)$$

where A is as in (11) and where

$$W = \int_1^{\rho_m} j^2 f(j)\,dj \qquad (15)$$

In Egghe (2003), it is generally proved that

$$\mu^* \geq \mu \qquad (16)$$

in all cases, a fact that will be reconfirmed by our practical calculations in the last section. Formula (16) means that, e.g. in the example of catalographic retrieval given above, that the cataloguer will, on the average $\left(\mu^{*}\right)$, encounter more books agreeing with a certain N-gram, than what could be expected from the average (μ) occurrence of this N-gram in the catalog.

This ends the general introduction of the informetric concepts and formulae needed in this paper. Since we will only work with N-grams and N-word phrases, all symbols f, g, P, A, T, μ, μ^{*} will have an index N in order to be able to distinguish between different values of $N = 1, 2, 3, 4, \ldots$.

As said above, in the sequel, all calculations will be exact (no approximations or simplifications). We will assume (8), (9) (i.e. the validity of Zipf's law) as explained in the introduction. We will also assume that letters occur independently in N-grams and that words occur independently in N-word phrases. Although this is not the case we assume this since, as shown in Egghe (2000b), we do not end up with analytical formulae for the rank-frequency distribution, if independence is not supposed. We trust that the formulae obtained in this paper describe the general N-tuple case to a large extent. The independence assumption can be mathematically formulated as

$$P\left(r_i \mid r_{i+1}, \ldots, r_N\right) = P_1\left(r_i\right) \tag{17}$$

i.e. the probability to have a letter or a word with rank r_i on the ith place $(i = 1, \ldots, N)$ is independent on ranks of the letters or the words on the places $i + 1, \ldots, N$, where the ranks refer to the single letter or single word case (i.e. $N = 1$, hence the notation $P_1\left(r_i\right)$) and where we put $P\left(r_N\right) = P_1\left(r_N\right)$.

III. The rank-frequency function of N-grams and N-word phrases

We can state and prove the following theorem.

Theorem III.1:
Let $N \in \mathbb{N}$ be fixed and assume (9) to be valid for $N = 1$ (and where we denote $P_1(r)$ for $P(r)$). Denote by $P_N(r)$ the rank-frequency probability density function of N-word phrases or N-grams. Then $r \in \left[0, T^N\right]$ and

$$P_N(r) = \frac{D^N}{\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^{\beta}} \tag{18}$$

where ξ_N^{-1} denotes the inverse of ξ_N and where ξ_N is the function

$$\xi_N(y) = \sum_{i=0}^{N-1} \frac{(-1)^{N+i-1} y \ln^i y}{i!} \tag{19}$$

and $\ln^i y = \underbrace{\ln(y)\ldots\ln(y)}_{i\ \text{times}}$, the ith power of $\ln(y)$.

Proof:
Since ranks are determined by (decreasing) productivity we have that $x = P_N(r)$, where

$$r = \operatorname{vol}\left\{(r_1,\ldots,r_N) \mid P(r_1,\ldots,r_N) \geq x\right\} \tag{20}$$

, where $P(r_1,\ldots,r_N)$ denotes the probability of occurrence of an N-gram or N-word phrase for which the ith letter (respectively word) has rank r_i (in the single occurrence), $i = 1,\ldots,N$. Here vol(S) denotes the volume of the N-dimensional set S. Now, by definition of conditional probability density (cf. Grimmett and Stirzaker (1985), p.61), repeatedly used:

$$
\begin{aligned}
&P(r_1,\ldots,r_N) \\
&\quad = P(r_1 \mid r_2,\ldots,r_N)P(r_2,\ldots,r_N) \\
&\quad = P(r_1 \mid r_2,\ldots,r_N)P(r_2 \mid r_3,\ldots,r_N)P(r_3,\ldots,r_N) \\
&\quad = \ \ldots \\
&\quad = P(r_1 \mid r_2,\ldots,r_N)P(r_2 \mid r_3,\ldots,r_N)\ldots P(r_{N-1} | r_N)P(r_N) \\
&\quad = P_1(r_1)P_1(r_2)\ldots P_1(r_{N-1})P_1(r_N)
\end{aligned}
$$

by (17). So, by (20) we have

$$r = \operatorname{vol}\{(r_1,\ldots,r_N) \mid P_1(r_1)P_1(r_2)\ldots P_1(r_N) \geq x\} \tag{21}$$

with $x = P_N(r)$, $x \in [0,1]$.

Note that, because of (8) and (9), the real ranks r_i should be lowered with 1 but, in (9), we can work with $r_i \in [1, T+1]$ itself and the set S is only a translation of the rank N-tuples $(r_1 - 1,\ldots,r_N - 1)$ over the vector $(1,\ldots,1)$ (N coordinates), so that the volume is the same. Hence we can use the r_is themselves in (21). Note, however, that r itself denotes the real rank of N-grams or N-word phrases. Indeed, let there be T letters (in case of N-grams) or T words (in case of N-word phrases (cf. (8)), then $r \in [0, T^N]$ and $r = T^N$ is obtained for $x = 0$, the set S being $S = [1, T+1]^N$ which volume is T^N and $r = 0$ is obtained for $x = \left(\frac{E}{A}\right)^N$ since then $\operatorname{vol}(S) = 0$ for the following reason: using (21) we have

$$P_1(r_1)P_1(r_2)\ldots P_1(r_N) \ge x = \left(\frac{E}{A}\right)^N$$

But by (7) and (8), each $P_1(r_i) \le \frac{E}{A}$. So

$$P_1(r_1)P_1(r_2)\ldots P_1(r_N) = \left(\frac{E}{A}\right)^N$$

But $0 \le P_1(r_i) \le \frac{E}{A}$ for every $i = 1,\ldots,N$ hence

$$P_1(r_i) = \frac{E}{A} \tag{22}$$

for every $i = 1,\ldots,N$. From (9) this implies

$$r_i = 1 \tag{23}$$

for every $i = 1,\ldots,N$. Hence $S = \{(1,1,\ldots,1)\}$, a singleton in $\mathbb{R}^N$ and hence $\mathrm{vol}(S) = 0$.

The inequality

$$P_1(r_1)P_1(r_2)\ldots P_1(r_N) \ge x$$

leads to, using (9)

$$\frac{D^N}{(r_1 r_2 \ldots r_N)^\beta} \ge x \tag{24}$$

hence

$$r_1 r_2 \ldots r_N \le \frac{D^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}} =: a \tag{25}$$

, by notation of a for reasons of simplicity. Formula (25) implies

$$1 \le r_1 \le \frac{a}{r_2 \ldots r_N} \tag{26}$$

This gives us the number of possible r_1s but dependent on the different r_2s$,\ldots,r_N$s that are possible. This will be determined now. Formula (26) yields

$$1 \le r_2 \le \frac{a}{r_3 \ldots r_N} \tag{27}$$

Formula (27) implies

$$1 \le r_3 \le \frac{a}{r_4 \ldots r_N} \tag{28}$$

and so on until

$$1 \le r_{N-1} \le \frac{a}{r_N} \tag{29}$$

and

$$1 \le r_N \le a \tag{30}$$

So $\mathrm{vol}(S)$ of (21) is found when we remark that r_1 ranges in an interval of length $\frac{a}{r_2 \ldots r_N} - 1$ (by (26)), where each $r_2,\ldots,r_N$ range as indicated in (27)-(30). Hence

$$r = a\int_{r_N=1}^{r_N=a}\frac{dr_N}{r_N}\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}\frac{dr_{N-1}}{r_{N-1}}\cdots\int_{r_2=1}^{r_2=\frac{a}{r_3\cdots r_N}}\frac{dr_2}{r_2}-\int_{r_N=1}^{r_N=a}dr_N\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}dr_{N-1}\cdots\int_{r_2=1}^{r_2=\frac{a}{r_3\cdots r_N}}dr_2 \qquad (31)$$

The evaluation of (31) is tedious but easy.

The first term in (31) (called (I)) is calculated as follows: since

$$\int_{r_2=1}^{r_2=\frac{a}{r_3\cdots r_N}}\frac{dr_2}{r_2} = -\ln\left(\frac{r_3\cdots r_N}{a}\right) \qquad (32)$$

(> 0 by (27)), we have that

$$(I) = -a\int_{r_N=1}^{r_N=a}\frac{dr_N}{r_N}\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}\frac{dr_{N-1}}{r_{N-1}}\cdots\int_{r_3=1}^{r_3=\frac{a}{r_4\cdots r_N}}\frac{\ln\left(\frac{r_3\cdots r_N}{a}\right)}{r_3}dr_3 \qquad (33)$$

But

$$\int_{r_3=1}^{r_3=\frac{a}{r_4\cdots r_N}}\frac{\ln\left(\frac{r_3\cdots r_N}{a}\right)}{r_3}dr_3 = -\frac{1}{2}\ln^2\left(\frac{r_4\cdots r_N}{a}\right)$$

as is readily seen. This value goes in (33) yielding

$$(I) = -a\int_{r_N=1}^{r_N=a}\frac{dr_N}{r_N}\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}\frac{dr_{N-1}}{r_{N-1}}\cdots\int_{r_4=1}^{r_4=\frac{a}{r_5\cdots r_N}}\frac{-\frac{1}{2}\ln^2\left(\frac{r_4\cdots r_N}{a}\right)}{r_4}dr_4 \qquad (34)$$

But

$$\int_{r_4=1}^{r_4=\frac{a}{r_5\cdots r_N}}\frac{-\frac{1}{2}\ln^2\left(\frac{r_4\cdots r_N}{a}\right)}{r_4}dr_4 = \frac{1}{3!}\ln^3\left(\frac{r_5\cdots r_N}{a}\right).$$

Note that each time the sign switches. This leads to

$$(I)= a\int_{r_N=1}^{r_N=a}\frac{dr_N}{r_N}\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}\frac{(-1)^{N-1}\ln^{N-3}\left(\frac{r_{N-1}r_N}{a}\right)}{(N-3)!\,r_{N-1}}dr_{N-1}$$

$$(I)= a\int_{r_N=1}^{r_N=a}\frac{dr_N}{r_N}\frac{(-1)^{N}\ln^{N-2}\left(\frac{r_N}{a}\right)}{(N-2)!}$$

$$(I)=\frac{(-1)^{N+1}a}{(N-1)!}\ln^{N-1}\left(\frac{1}{a}\right)$$

$$(I)=\frac{a\ln^{N-1}a}{(N-1)!}\geq 0 \qquad (35)$$

since $a=\frac{D^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}}\geq r_1\ldots r_N\geq 1$, using (25).

Now we calculate the second term in (31), called (II).

$$(II)= -\int_{r_N=1}^{r_N=a}dr_N\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}dr_{N-1}\cdots\int_{r_2=1}^{r_2=\frac{a}{r_3\ldots r_N}}dr_2$$

$$(II)= -\int_{r_N=1}^{r_N=a}dr_N\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}dr_{N-1}\cdots\int_{r_3=1}^{r_3=\frac{a}{r_4\ldots r_N}}\left(\frac{a}{r_3\ldots r_N}-1\right)dr_3$$

$$(II)= -\int_{r_N=1}^{r_N=a}dr_N\int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}}dr_{N-1}\cdots\int_{r_4=1}^{r_4=\frac{a}{r_5\ldots r_N}}\left[\frac{a}{r_4\ldots r_N}\ln\left(\frac{a}{r_4\ldots r_N}\right)-\frac{a}{r_4\ldots r_N}+1\right]dr_4$$

$$(\mathrm{II}) = -\int_{r_N=1}^{r_N=a} dr_N \int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}} dr_{N-1} \;\ldots \int_{r_5=1}^{r_5=\frac{a}{r_6\ldots r_N}} \left[\frac{1}{2}\frac{a}{r_5\ldots r_N}\ln^2\left(\frac{a}{r_5\ldots r_N}\right) - \frac{a}{r_5\ldots r_N}\ln\left(\frac{a}{r_5\ldots r_N}\right) + \frac{a}{r_5\ldots r_N} - 1\right] dr_5$$

$$(\mathrm{II}) = -\int_{r_N=1}^{r_N=a} dr_N \int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}} dr_{N-1} \;\ldots$$

$$\int_{r_6=1}^{r_6=\frac{a}{r_7\ldots r_N}} \left[\frac{1}{3!}\frac{a}{r_6\ldots r_N}\ln^3\left(\frac{r_6\ldots r_N}{a}\right) - \frac{1}{2}\frac{a}{r_6\ldots r_N}\ln^2\left(\frac{r_6\ldots r_N}{a}\right) - \frac{a}{r_6\ldots r_N}\ln\left(\frac{r_6\ldots r_N}{a}\right) - \frac{a}{r_6\ldots r_N} + 1\right] dr_6$$

In general we have $(j+3 = 2,\ldots,N)$

$$(\mathrm{II}) = -\int_{r_N=1}^{r_N=a} dr_N \;\ldots \int_{r_{j+3}=1}^{r_{j+3}=\frac{a}{r_{j+4}\ldots r_N}} \left[(-1)^j \frac{a}{r_{j+3}\ldots r_N}\sum_{i=1}^{j}\frac{1}{i!}\ln^i\left(\frac{r_{j+3}\ldots r_N}{a}\right) + \frac{a}{r_{j+3}\ldots r_N}(-1)^j + (-1)^{j-1}\right] dr_{j+3}$$

Hence

$$(\mathrm{II}) = -\int_{r_N=1}^{r_N=a}\left[(-1)^{N-3}\frac{a}{r_N}\sum_{i=1}^{N-3}\frac{1}{i!}\ln^i\left(\frac{r_N}{a}\right) + (-1)^{N-4}\right] dr_N + \int_{r_N=1}^{r_N=a}\frac{a}{r_N}(-1)^{N-3}\, dr_N$$

$$(\mathrm{II}) = (-1)^{N-2}\, a\sum_{i=1}^{N-3}\frac{1}{(i+1)!}\left(-\ln^{i+1}\left(\frac{1}{a}\right)\right) - (-1)^{N-4}(a-1) - (-1)^{N-3}\, a\ln a$$

$$(\mathrm{II}) = (-1)^{N-1}\sum_{i=1}^{N-3}\frac{a(-1)^{i+1}\ln^{i+1} a}{(i+1)!} + (-1)^N a\ln a + (-1)^{N-1} a + (-1)^N$$

$$(\mathrm{II}) = (-1)^{N-1}\sum_{i=0}^{N-2}\frac{(-1)^i a\ln^i a}{i!} + (-1)^N$$

$$(II) = \sum_{i=0}^{N-2} \frac{(-1)^{N+i-1} a \ln^i a}{i!} + (-1)^N \qquad (36)$$

Now (35) and (36) yield, by (31)

$$r = \frac{a \ln^{N-1} a}{(N-1)!} + \sum_{i=0}^{N-2} \frac{(-1)^{N+i-1} a \ln^i a}{i!} + (-1)^N$$

$$r = \sum_{i=0}^{N-1} \frac{(-1)^{N+i-1} a \ln^i a}{i!} + (-1)^N \qquad (37)$$

Using (25) and the fact that $x = P_N(r)$, we have by (37)

$$r + (-1)^{N-1} = \xi_N\left(\frac{D^{\frac{N}{\beta}}}{(P_N(r))^{\frac{1}{\beta}}}\right), \qquad (38)$$

where

$$\xi_N(y) = \sum_{i=0}^{N-1} \frac{(-1)^{N+i-1} y \ln^i y}{i!}$$

, i.e. formula (19). By (25) the arguments of the logarithms, appearing in ξ_N are greater than or equal to 1, hence positive. Note that ξ_N is an injection on $[1,+\infty[$. Indeed:

$$\xi_N'(y) = \sum_{i=0}^{N-1} \frac{(-1)^{N+i-1} \ln^i y}{i!} + \sum_{i=1}^{N-1} \frac{(-1)^{N+i-1} \ln^{i-1} y}{(i-1)!}$$

$$\xi_N'(y) = \sum_{i=0}^{N-1} \frac{(-1)^{N+i-1} \ln^i y}{i!} + \sum_{i=0}^{N-2} \frac{(-1)^{N+i} \ln^i y}{i!}$$

$$\xi_N^{'}(y) = \frac{\ln^{N-1} y}{(N-1)!} > 0 \tag{39}$$

on $y \in]1,+\infty[$. So ξ_N is a strictly increasing function on $[1,+\infty[$ and hence an injection. But

$$a = \frac{D^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}} \geq 1$$

by (25), hence we can take the inverse of ξ_N in (38) yielding

$$P_N(r) = \frac{D^N}{\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^{\beta}}$$

where ξ_N^{-1} denotes the inverse of the function ξ_N. ~

The function $P_N(r)$ is not simple. We have the following corollary, proved in Egghe (1999) and Egghe (2000a) as an approximate result:

Corollary III.2:

If r is large, we have that

$$P_N(r) \approx \frac{D^N}{\left(\chi_N^{-1}((N-1)!\,r)\right)^{\beta}} \tag{40}$$

where χ_N^{-1} is the inverse of the function

$$\chi_N(y) = y \ln^{N-1}(y) \tag{41}$$

(again $\ln^{N-1}(y)$ denotes the (N – 1)th power of ln(y)).

Proof:

The number r large enough forces all the ranks $r_1,...,r_N$ to be large by (21). Since r_1 is large we have by (26) that

$$\frac{a}{r_2...r_N} - 1 \approx \frac{a}{r_2...r_N}.$$

In other words, in the proof of the above theorem we only calculate (I) for r and put $(II) \approx 0$. By (35), this yields the result. ~

This approximation was used in Egghe (1999, 2000a) because evaluating (II) did not seem to lead to any useful result. Indeed, formulae (18) and (19) are much more complicated than (40) and (41) and if it were not for the results in the sequel we would not consider these intricate results as important. We are, however, lucky: in the next subsection we will derive the size-frequency function f_N linked to the above rank-frequency distribution P_N and we will show that the exact result (18) leads to a very simple formula for f_N, simpler than the one derived from the inexact (40)!

The derivation of the size-frequency function f_N is based on the general formulae of Section II on the link between the rank- and the size-frequency function. Therefore we first have to determine the rank-frequency function (called g in Section II and called g_N here to show the N-dependence) derived from the rank-frequency density function P_N in (18). g_N follows from P_N by (7), i.e. simply by multiplying with the total number of items in the case of N-grams or N-word phrases, which we will denote by A_N (in Section II this is denoted by A). Consequently, we have

$$g_N(r) = \frac{A_N D^N}{\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^{\beta}} \qquad (42)$$

for $r \in \left[0, T^N\right]$, using Theorem III.1.

In the proof of Theorem III.1 we showed that ξ_N strictly increases, hence the same is true for ξ_N^{-1}, so g_N strictly decreases, using (42). From (39) it follows that $\xi_N'(y) > 0$ and $\xi_N''(y) > 0$ on $]1,+\infty[$. This can be used in (42) to show that g_N is convexly decreasing, as it should (by the very definition of P_N). We leave this as an exercise.

There are not many practical data on N-grams or N-word phrases. A convexly decreasing rank-frequency function for N-grams can be found in Cavnar and Trenkle (1994). These authors use the name “Zipfian” distribution which, visually, and probably also statistically, is a normal observation. In this Section III we only tried to show the mathematical link between 1-gram (1-word phrase)-theory (i.e. Lotkaian, Zipfian informetrics) and N-gram (N-word phrase)-theory. In general, the above theory (and the one to follow on the size-frequency function) can be considered as the mathematical theory on how to describe informetrically the Cartesian product of N IPPs with the same Zipfian rank-frequency distribution.

The result (42) on g_N is intricate and not easy to work with. In the next section we will determine the size-frequency function f_N that is equivalent with the rank-frequency function g_N, using the model in Section II. The result on f_N will be surprisingly simple (although its derivation is, once more, tedious).

IV. The size-frequency function of N-grams and N-word phrases derived from Section III

We have the following theorem.

Theorem IV.1:

The size-frequency function f_N that is equivalent with the rank-frequency function g_N of (42) is given by

$$f_N(j) = \frac{C}{j^{1+\frac{1}{\beta}}} \ln^{N-1}\left(\frac{\rho_m(N)}{j}\right) \qquad (43)$$

for $j \in \left[1, \rho_m(N)\right]$, where $\rho_m(N)$ is the maximal item density in the case if N-grams or N-word phrases, given by

$$\rho_m(N) = A_N D^N \tag{44}$$

and where C is the constant

$$C = \frac{\rho_m(N)^{\frac{1}{\beta}}}{\beta^N (N-1)!}. \tag{45}$$

Proof:

By the very definition of size-frequency function, we have (see formula (5)):

$$f_N(j) = -\frac{1}{g'_N\left(g_N^{-1}(j)\right)} \tag{46}$$

for $j \in \left[1, \rho_m(N)\right]$ with $\rho_m(N)$ the maximal item density in the case of N-grams or N-word phrases. Formula (42) yields

$$g_N(r)\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^{\beta} = A_N D^N$$

Hence, taking derivatives

$$g'_N(r)\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^{\beta} + g_N(r)\beta\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^{\beta-1} . \frac{d\left(\xi_N^{-1}\right)}{dx}\left(x = r + (-1)^{N-1}\right) = 0$$

where

$$\frac{d\left(\xi_N^{-1}\right)}{dx}\left(x = r + (-1)^{N-1}\right)$$

means: the derivative of the function ξ_N^{-1} in the point $r+(-1)^{N-1}$. So

$$g_N'(r)\xi_N^{-1}\left(r+(-1)^{N-1}\right)=\frac{-\beta g_N(r)}{\xi_N'\left(\xi_N^{-1}\left(r+(-1)^{N-1}\right)\right)}. \tag{47}$$

But, by (39)

$$g_N'(r)\xi_N^{-1}\left(r+(-1)^{N-1}\right)=\frac{-\beta g_N(r)}{\frac{1}{(N-1)!}\ln^{N-1}\left(\xi_N^{-1}\left(r+(-1)^{N-1}\right)\right)} \tag{48}$$

Now we use (42), yielding

$$g_N'(r)\xi_N^{-1}\left(r+(-1)^{N-1}\right)=\frac{-\beta A_N D^N}{\frac{\left(\xi_N^{-1}\left(r+(-1)^{N-1}\right)\right)^{\beta}}{(N-1)!}\ln^{N-1}\left(\xi_N^{-1}\left(r+(-1)^{N-1}\right)\right)} \tag{49}$$

So

$$g_N'(r)=\frac{-\beta A_N D^N}{\frac{\left(\xi_N^{-1}\left(r+(-1)^{N-1}\right)\right)^{\beta+1}}{(N-1)!}\ln^{N-1}\left(\xi_N^{-1}\left(r+(-1)^{N-1}\right)\right)} \tag{50}$$

Since $j=g_N(r)$ denotes the item density (by definition (4)), we have by (42) that

$$\xi_N^{-1}\left(r+(-1)^{N-1}\right)=\left(\frac{A_N D^N}{j}\right)^{\frac{1}{\beta}} \tag{51}$$

in the point $r = g_N^{-1}(j)$. So (51) in (50) yields

$$g_N'\left(g_N^{-1}(j)\right) = \frac{-\beta A_N D^N}{\left(\frac{A_N D^N}{j}\right)^{1+\frac{1}{\beta}} \frac{\ln^{N-1}\left(\left(\frac{A_N D^N}{j}\right)^{\frac{1}{\beta}}\right)}{(N-1)!}} \qquad (52)$$

which yields, by (46) the result

$$f_N(j) = \frac{\left(A_N D^N\right)^{\frac{1}{\beta}}}{\beta^N j^{1+\frac{1}{\beta}}(N-1)!}\ln^{N-1}\left(\frac{A_N D^N}{j}\right) \qquad (53)$$

, $j \in \left[1, \rho_m(N)\right]$, a remarkably simple result. By definition of $\rho_m(N)$ and g_N we have

$$\rho_m(N) = g_N(0) = \frac{A_N D^N}{\left(\xi_N^{-1}\left((-1)^{N-1}\right)\right)^{\beta}} \qquad (54)$$

by (42). But $\xi_N(1) = (-1)^{N-1}$ as follows readily from (19). Hence, since we showed in Theorem III.1 that ξ_N is an injection on $[1,+\infty[$, we have that $\xi_N^{-1}\left((-1)^{N-1}\right) = 1$ and so, from (54)

$$\rho_m(N) = A_N D^N$$

proving (44). Now (53) and (54) give

$$f_N(j) = \frac{C}{j^{1+\frac{1}{\beta}}}\ln^{N-1}\left(\frac{\rho_m(N)}{j}\right)$$

with C as in (45), hence we have proved (43), for $j \in \left[1, \rho_m(N)\right]$. ~

Note that, in terms of Lotka's α, see (3), we have that (43) also reads as

$$f_N(j) = \frac{C}{j^{\alpha}} \ln^{N-1}\left(\frac{\rho_m(N)}{j}\right) \qquad (55)$$

hence a product of a power law and a power of a logarithm. It is easy to see that $f_N^{'} < 0$ and $f_N^{''} > 0$ hence f_N is convexly decreasing on $\left[1, \rho_m(N)\right] = \left[1, A_N D^N\right]$.

Note also that g_N and f_N, for N = 1, reduce to the given laws of Zipf and Lotka (as it should). Indeed, for f_1 this is clear (with $C = \frac{\rho_m^{\frac{1}{\beta}}}{\beta}$ as follows from (45), agreeing with the results in Rousseau (1990), since we supposed Zipf's law for g_1). For g_1, we have by (42)

$$g_1(r) = \frac{AD}{\left(\xi_1^{-1}(r+1)\right)^{\beta}}$$

$$= \frac{AD}{(r+1)^{\beta}}$$

since $\xi_1(y) = y$ by (19), and hence

$$g_1(r) = \frac{E}{(r+1)^{\beta}},$$

the same function as (8), using that we denoted $D = \frac{E}{A}$.

In the next section we will use the size-frequency function f_N to calculate the averages μ (here denoted as μ_N) of items per source and μ^* (here denoted as μ_N^*) being the Type/Token-Taken average as discussed in Section II. In terms of the present notations we could say that the Type/Token-Taken theory of Section II was based on f_1; in the next section we will use f_N $(N \geq 2)$. Of course, the general defining formulae for μ and μ^* (i.e. for general size-frequency functions) of Section II also apply here.

V. Type/Token averages μ_N and Type/Token-Taken averages μ_N^* for N-grams and N-word phrases

As follows from formulae (11), (12), (14) and (15), we have that the TT average μ_N and the TTT averages μ_N^* are given by

$$\mu_N = \frac{A_N}{T^N} \qquad (56)$$

$$\mu_N^* = \frac{W_N}{A_N} \qquad (57)$$

where

$$T^N = \int_1^{\rho_m(N)} f_N(j)dj \qquad (58)$$

$$A_N = \int_1^{\rho_m(N)} jf_N(j)dj \qquad (59)$$

$$W_N = \int_1^{\rho_m(N)} j^2 f_N(j)dj \qquad (60)$$

and where f_N is given by (43). All these integrals are tedious to calculate but we can use the following formula found in Gradshteyn and Ryzhik (1965) (p.203 (2.722)):

$$\int x^n \ln^m x \quad dx = \frac{x^{n+1}}{m+1}\sum_{k=0}^{m}(-1)^k (m+1)m(m-1)\ldots(m-k+1)\frac{\ln^{m-k} x}{(n+1)^{k+1}} \tag{61}$$

valid for all $n \in \mathbb{R} \setminus \{-1\}$ and $m \in \mathbb{N}$.

For the calculation of T^N (i.e. in function of $\rho_m(N)$, which will be our free parameter, just as it was the case with ρ_m in Egghe (2003)), we have two equivalent alternatives: or we can calculate (58) directly or (which we will do here) use the following short argument. We note that $j = g_N(r)$ and hence $1 = g_N(T^N)$ ($r = T^N$ was the highest rank as proved in Theorem III.1). Formula (42) yields

$$1 = \frac{A_N D^N}{\left(\xi_N^{-1}\left(T^N + (-1)^{N-1}\right)\right)^{\beta}}$$

so

$$T^N + (-1)^N = \xi_N\left(\left(A_N D^N\right)^{\frac{1}{\beta}}\right)$$

Using (19) we have

$$T^N + (-1)^{N-1} = \sum_{i=0}^{N-1}\frac{(-1)^{N+i-1}\left(A_N D^N\right)^{\frac{1}{\beta}}\ln^i\left(\left(A_N D^N\right)^{\frac{1}{\beta}}\right)}{i!}$$

hence, by (44)

$$T^N = (-1)^N + \sum_{i=0}^{N-1}\frac{(-1)^{N+i-1}\left(\rho_m(N)\right)^{\frac{1}{\beta}}\ln^i\left(\left(\rho_m(N)\right)^{\frac{1}{\beta}}\right)}{i!} \tag{62}$$

valid for all $N \in \mathbb{N}$ and all $\beta > 0$.

We are left with the calculation of (59) and (60), using (43). We have

$$A_N = \int_1^{\rho_m(N)} \frac{C}{\varphi^{\frac{1}{\beta}}} \ln^{N-1}\left(\frac{\rho_m(N)}{\varphi}\right) d\varphi . \qquad (63)$$

Since

$$d\left(\frac{\rho_m(N)}{\varphi}\right) = -\frac{\rho_m(N)}{\varphi^2} d\varphi$$

we have that

$$\int \frac{\ln^{N-1}\left(\frac{\rho_m(N)}{\varphi}\right)}{\varphi^{\frac{1}{\beta}}} d\varphi = -\rho_m(N)^{1-\frac{1}{\beta}} \int \left(\frac{\rho_m(N)}{\varphi}\right)^{\frac{1}{\beta}-2} \ln^{N-1}\left(\frac{\rho_m(N)}{\varphi}\right) d\left(\frac{\rho_m(N)}{\varphi}\right) \qquad (64)$$

So, for $\beta > 0$, $\beta \neq 1$ we can apply (61) yielding

$$\int \frac{\ln^{N-1}\left(\frac{\rho_m(N)}{\varphi}\right)}{\varphi^{\frac{1}{\beta}}} d\varphi = \frac{-1}{N\varphi^{\frac{1}{\beta}-1}} \sum_{k=0}^{N-1} (-1)^k N(N-1)\ldots(N-k) \frac{\ln^{N-k-1}\left(\frac{\rho_m(N)}{\varphi}\right)}{\left(\frac{1}{\beta}-1\right)^{k+1}}$$

$$= \sum_{k=1}^{N} \frac{(-1)^k}{\varphi^{\frac{1}{\beta}-1}} (N-1)(N-2)\ldots(N-k+1) \frac{\ln^{N-k}\left(\frac{\rho_m(N)}{\varphi}\right)}{\left(\frac{1}{\beta}-1\right)^{k}}$$

where we note that, for $k = 1$, we have to take $(N-1)(N-2)\ldots(N-k+1) = 1$. (63) now yields, using (45)

$A_N =$

$$\frac{(\rho_m(N))^{\frac{1}{\beta}}}{\beta^N (N-1)!}\left[\frac{(-1)^N (N-1)!}{(\rho_m(N))^{\frac{1}{\beta}-1}\left(\frac{1}{\beta}-1\right)^N} - \sum_{k=1}^{N}\frac{(-1)^k (N-1)\ldots(N-k+1)\ln^{N-k}(\rho_m(N))}{\left(\frac{1}{\beta}-1\right)^k}\right] \qquad (65)$$

valid for all N and $\beta > 0$, $\beta \neq 1$ and noting that, for $k = 1$, $(N-1)\ldots(N-k+1) = 1$.

For $\beta = 1$, we have

$$A_N = \int_1^{\rho_m(N)} \frac{C}{j}\ln^{N-1}\left(\frac{\rho_m(N)}{j}\right)dj$$

$$A_N = \frac{\rho_m(N)\ln^N(\rho_m(N))}{N!} \qquad (66)$$

as is easily calculated using (63) and (45) for $\beta = 1$.

For W_N we have

$$W_N = \int_1^{\rho_m(N)} \frac{C}{j^{\frac{1}{\beta}-1}}\ln^{N-1}\left(\frac{\rho_m(N)}{j}\right)dj \qquad (67)$$

But, using (63), we have

$$\int \frac{\ln^{N-1}\left(\frac{\rho_m(N)}{j}\right)}{j^{\frac{1}{\beta}-1}}dj = -(\rho_m(N))^{2-\frac{1}{\beta}}\int\left(\frac{\rho_m(N)}{j}\right)^{\frac{1}{\beta}-3}\ln^{N-1}\left(\frac{\rho_m(N)}{j}\right)d\left(\frac{\rho_m(N)}{j}\right) \qquad (68)$$

which can be calculated, using (61) for all $\beta \neq \frac{1}{2}$. This gives

$$\int \frac{\ln^{N-1}\left(\frac{\rho_m(N)}{\varphi}\right)}{\varphi^{\frac{1}{\beta}-1}}d\varphi = \frac{-1}{\varphi^{\frac{1}{\beta}-2}N}\sum_{k=0}^{N-1}(-1)^k N(N-1)\dots(N-k)\frac{\ln^{N-k-1}\left(\frac{\rho_m(N)}{\varphi}\right)}{\left(\frac{1}{\beta}-2\right)^{k+1}}$$

$$=\sum_{k=1}^{N}\frac{(-1)^k}{\varphi^{\frac{1}{\beta}-2}}(N-1)\dots(N-k+1)\frac{\ln^{N-k}\left(\frac{\rho_m(N)}{\varphi}\right)}{\left(\frac{1}{\beta}-2\right)^{k}}$$

where, for $k=1$, we have to take $(N-1)\dots(N-k+1)=1$. Hence we have, from (67), using (45)

$W_N =$

$$\frac{(\rho_m(N))^{\frac{1}{\beta}}}{\beta^N(N-1)!}\left[\frac{(-1)^N(N-1)!}{(\rho_m(N))^{\frac{1}{\beta}-2}\left(\frac{1}{\beta}-2\right)^N}-\sum_{k=1}^{N}\frac{(-1)^k(N-1)\dots(N-k+1)\ln^{N-k}(\rho_m(N))}{\left(\frac{1}{\beta}-2\right)^k}\right] \qquad (69)$$

valid for all N and $\beta \neq \frac{1}{2}$ and where we have to take $(N-1)\dots(N-k+1)=1$ for $k=1$.

For $\beta=\frac{1}{2}$ we have, using (63)

$$\int \frac{\ln^{N-1}\left(\frac{\rho_m(N)}{\varphi}\right)}{\varphi}d\varphi = -\frac{\ln^{N}\left(\frac{\rho_m(N)}{\varphi}\right)}{N}$$

So (67) and (45) yield

$$W_N = \frac{2^N (\rho_m(N))^2}{N!} \ln^N (\rho_m(N)) \tag{70}$$

for all N and $\beta = \frac{1}{2}$.

With these formulae for T^N, A_N and W_N we are able to calculate μ_N and μ_N^* via (56) and (57). We will also compare these values with the corresponding values of μ_1 and μ_1^*, i.e. TT and TTT averages in the case of 1-grams (single letters) or of 1-word phrases (single words) as developed in Egghe (2003).

As examples we will take $\beta = 1$ (i.e. Lotka's $\alpha = 2$) and $\beta = \frac{1}{2}$ (i.e. Lotka's $\alpha = 3$) and we will take N=1, 2, 3: the case of 2(3)-grams or 2(3)-word phrases in comparison with single letters or words will be informative enough for higher values of N. In addition the cases $N = 2$ and $N = 3$ are the most important cases for all applications.

Let us take $\beta = 1$ first. For $N = 2$ we have from (62), (66) and (69)

$$T^2 = 1 - \rho_m(2) + \rho_m(2)\ln(\rho_m(2)) \tag{71}$$

$$A_2 = \frac{1}{2}\rho_m(2)\ln^2(\rho_m(2)) \tag{72}$$

$$W_2 = (\rho_m(2))^2 - \rho_m(2)\ln(\rho_m(2)) - \rho_m(2) \tag{73}$$

Hence

$$\mu_2 = \frac{\rho_m(2)\ln^2(\rho_m(2))}{2(1 - \rho_m(2) + \rho_m(2)\ln(\rho_m(2)))} \tag{74}$$

$$\mu_2^* = \frac{2\left(\rho_m(2) - \ln\left(\rho_m(2)\right) - 1\right)}{\ln^2\left(\rho_m(2)\right)} \qquad (75)$$

which yields Table 1.

Table 1. Values of μ_2 and μ_2^* for diverse values of $\rho_m(2)$, for $\beta = 1$

$\rho_m(2)$	1.5	2	3	5	10	100
μ_2	1.140	1.244	1.397	1.600	1.890	2.933
μ_2^*	1.150	1.277	1.494	1.846	2.526	8.902

This can be compared with the values of μ_1 and μ_1^*, i.e. the non-composed case. For $\beta = 1$ (hence $\alpha = 2$) we use the formulae (cf. Egghe (2003))

$$\mu = \mu_1 = \frac{\ln \rho_m}{1 - \frac{1}{\rho_m}} \qquad (76)$$

and

$$\mu^* = \mu_1^* = \frac{\rho_m - 1}{\ln \rho_m} \qquad (77)$$

yielding Table 2.

Table 2. Values of μ_1 and μ_1^* for diverse values of ρ_m, for $\beta = 1$

ρ_m	1.5	2	3	5	10	100
μ_1	1.216	1.386	1.648	2.012	2.558	4.652
μ_1^*	1.233	1.443	1.820	2.485	3.909	21.498

We see that, for the same value of the input “seed “ $\rho_m(2)$ or ρ_m, we have that the values μ_1 and μ_1^* are larger than the values μ_2 and μ_2^* respectively. We also see that $\mu_2^* - \mu_2 < \mu_1^* - \mu_1$ showing that the average screen lengths (e.g. in the case of the use of 2-grams by a cataloger) are shorter than the ones given in the 1-gram case. Note further that $\mu_1^* > \mu_1$ and $\mu_2^* > \mu_2$ as it should, following (16).

Now we calculate the case $N = 3$, still with $\beta = 1$. We have from (62), (66) and (69)

$$T^3 = -1 + \rho_m(3) - \rho_m(3)\ln(\rho_m(3)) + \frac{1}{2}\rho_m(3)\ln^2(\rho_m(3)) \tag{78}$$

$$A_3 = \frac{1}{6}\rho_m(3)\ln^3(\rho_m(3)) \tag{79}$$

$$W_3 = (\rho_m(3))^2 - \frac{1}{2}\rho_m(3)\ln^2(\rho_m(3)) - \rho_m(3)\ln(\rho_m(3)) - \rho_m(3) \tag{80}$$

Hence

$$\mu_3 = \frac{\rho_m(3)\ln^3(\rho_m(3))}{-6 + 6\rho_m(3) - 6\rho_m(3)\ln(\rho_m(3)) + 3\rho_m(3)\ln^2(\rho_m(3))} \tag{81}$$

$$\mu_3^* = \frac{6\rho_m(3) - 3\ln^2(\rho_m(3)) - 6\ln(\rho_m(3)) - 6}{\ln^3(\rho_m(3))} \tag{82}$$

which yields Table 3.

Table 3. Values of μ_3 and μ_3^* for diverse

values of $\rho_m(3)$, for $\beta = 1$

$\rho_m(3)$	1.5	2	3	5	10	100
μ_3	1.103	1.179	1.288	1.431	1.630	2.329
μ_3^*	1.110	1.200	1.348	1.577	1.989	5.148

The same comments as for μ_2, μ_2^*, given above, can be given here for μ_3, μ_3^*. Note again that the values of μ_3, μ_3^* are smaller than the values of μ_2, μ_2^* respectively.

We, finally, give formulae for $\beta = \frac{1}{2}$ and N = 2, 3 and compare with the case $N = 1$. For $N = 2$ and $\beta = \frac{1}{2}$ we have the following formulae, following from (62), (65) and (67)

$$T^2 = 1 - (\rho_m(2))^2 + 2(\rho_m(2))^2 \ln(\rho_m(2)) \tag{83}$$

$$A_2 = 4\rho_m(2) + 4(\rho_m(2))^2 \ln(\rho_m(2)) - 4(\rho_m(2))^2 \tag{84}$$

$$W_2 = 2(\rho_m(2))^2 \ln^2(\rho_m(2)) \tag{85}$$

Hence we have

$$\mu_2 = \frac{4\rho_m(2) + 4(\rho_m(2))^2 \ln(\rho_m(2)) - 4(\rho_m(2))^2}{1 - (\rho_m(2))^2 + 2(\rho_m(2))^2 \ln(\rho_m(2))} \tag{86}$$

$$\mu_2^* = \frac{\rho_m(2)\ln^2(\rho_m(2))}{2 + 2\rho_m(2)\ln(\rho_m(2)) - 2\rho_m(2)} \tag{87}$$

yielding Table 4.

Table 4. Values of μ_2 and μ_2^* for diverse

values of $\rho_m(2)$, for $\beta = \frac{1}{2}$

$\rho_m(2)$	1.5	2	3	5	10	100
μ_2	1.130	1.214	1.321	1.433	1.552	1.761
μ_2^*	1.140	1.244	1.397	1.600	1.890	2.933

Compare now with the case $N = 1$, $\beta = \frac{1}{2}$ (hence $\alpha = 3$), using the formulae (cf. Egghe (2003)):

$$\mu = \mu_1 = \frac{2\rho_m}{\rho_m + 1} \tag{88}$$

$$\mu^* = \mu_1^* = \frac{\ln \rho_m}{1 - \frac{1}{\rho_m}} \tag{89}$$

yielding Table 5.

Table 5. Values of μ_1 and μ_1^* for diverse

values of ρ_m, for $\beta = \frac{1}{2}$

ρ_m	1.5	2	3	5	10	100
μ_1	1.200	1.333	1.500	1.667	1.818	1.980
μ_1^*	1.216	1.386	1.648	2.012	2.558	4.652

For $N = 3$, $\beta = \frac{1}{2}$ we have now, using (62), (65) and (67)

$$T^3 = -1 + (\rho_m(3))^2 - 2(\rho_m(3))^2 \ln(\rho_m(3)) + 2(\rho_m(3))^2 \ln^2(\rho_m(3)) \qquad (90)$$

$$A_3 = -8\rho_m(3) + 4(\rho_m(3))^2 \ln^2(\rho_m(3)) - 8(\rho_m(3))^2 \ln(\rho_m(3)) + 8(\rho_m(3))^2 \qquad (91)$$

$$W_3 = \frac{4}{3}(\rho_m(3))^2 \ln^3(\rho_m(3)) \qquad (92)$$

Hence

$$\mu_3 = \frac{-8\rho_m(3) + 4(\rho_m(3))^2 \ln^2(\rho_m(3)) - 8(\rho_m(3))^2 \ln(\rho_m(3)) + 8(\rho_m(3))^2}{-1 + (\rho_m(3))^2 - 2(\rho_m(3))^2 \ln(\rho_m(3)) + 2(\rho_m(3))^2 \ln^2(\rho_m(3))} \qquad (93)$$

$$\mu_3^* = \frac{\rho_m(3)\ln^3(\rho_m(3))}{-6 + 3\rho_m(3)\ln^2(\rho_m(3)) - 6(\rho_m(3))\ln(\rho_m(3)) + 6(\rho_m(3))} \qquad (94)$$

yielding Table 6.

Table 6. Values of μ_3 and μ_3^* for diverse

values of $\rho_m(3)$, for $\beta = \frac{1}{2}$

$\rho_m(3)$	1.5	2	3	5	10	100
μ_3	1.097	1.160	1.241	1.330	1.429	1.635
μ_3^*	1.103	1.179	1.288	1.431	1.630	2.329

We see again that the same tendencies of the comparison of μ_1, μ_1^*, μ_2, μ_2^*, μ_3, μ_3^* are found as in the case $\beta = 1$.

We close with an open problem.

Open Problem:

Describe the TT average and TTT average in case of N-grams where the number of items is limited to the number of documents in a database (e.g. an OPAC, used by a cataloger, as described in Section II). Since, here, the number of items (denoted A) is fixed and since there are T^N N-grams (cf. Theorem III.1), we might end up, for not even very large N with the relation $T^N > A$, hence with more sources than items, which is out of the scope of the informetric theory which was briefly described in Section II.

References

R.H. Baayen (2001). Word Frequency Distributions. Kluwer Academic Publishers, Dordrecht, the Netherlands, 2001.

W.B. Cavnar and J.M. Trenkle (1994). N-gram-based text categorization. IN: Proceedings of the third Annual Symposium on Document Analysis and Information Retrieval, 161-175, University of Las Vegas, USA, 1994.

J.D. Cohen (1995). Highlights: language- and domain-independent automatic indexing terms for abstracting. Journal of the American Society for Information Science 46(3), 162-174, 1995.

M. Damashek (1995). Gauging similarity with N-grams: language-independent categorization of text. Science 267 (10 February 1995), 843-848, 1995.

L. Egghe (1989). The Duality of Informetric Systems with Applications to the empirical Laws. Ph. D. Thesis, City University, London (UK), 1989.

L. Egghe (1990). The duality of informetric systems with applications to the empirical laws. Journal of Information Science 16(1), 17-27, 1990.

L. Egghe (1999). On the law of Zipf-Mandelbrot for multi-word phrases. Journal of the American Society for Information Science 50(3), 233-241, 1999.

L. Egghe (2000a). The distribution of N-grams. Scientometrics 47(2), 237-252, 2000.

L. Egghe (2000b). General study of the distribution of N-tuples of letters or words based on the distributions of the single letters or words. Mathematical and Computer Modelling 31, 35-41, 2000.

L. Egghe (2003). Type/Token-Taken informetrics. Journal of the American Society for Information Science and Technology 54(7), 603-610, 2003.

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

I.S. Gradshteyn and I.M. Ryzhik (1965). Table of Integrals, Series and Products. Academic Press, New York, USA, 1965.

G.R. Grimmett and D.R. Stirzaker (1985). Probability and random Processes. Clarendon Press, Oxford, UK, 1985.

D.A. Grossman and O. Frieder (1998). Information Retrieval. Algorithms and Heuristics. Kluwer Academic Pubishers, Dordrecht, the Netherlands, 1998.

G. Herdan (1960). Type-Token Mathematics. A Textbook of mathematical Linguistics. Mouton, ‘s Gravenhage, the Netherlands, 1960.

M. Nelson and J.S. Downie (2001). Informetric analysis of a music database: distributions of intervals. Proceedings of the eighth International Conference on Scientometrics and Informetrics (M. Davis and C.S. Wilson, eds.), 477-484, BIRG, University of New South Wales, Sydney, Australia, 2001.

A.M. Robertson and P. Willett (1998). Application of N-grams in textual information systems. Journal of Documentation 54(1), 48-69, 1998.

R. Rousseau (1990). Relations between continuous versions of bibliometric laws. Journal of the American Society for Information Science 41(3), 197-203, 1990.

E.J. Yannakoudakis, I. Tsomokos and P.J. Hutton (1990). N-grams and their implication to natural language understanding. Pattern Recognition 23(5), 509-528, 1990.

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