The exact rank-frequency function and size-frequency function of N-grams and N-word phrases with applications Link
Peer-reviewed author version

Made available by Hasselt University Library in Document Server@UHasselt

Reference (Published version):
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
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
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 This leads to an intricate rank-frequency function for N-grams and N-word phrases (as we knew before from unpublished calculations) but leads surprisingly, to a very simple size-frequency function \( f_N \) for N-grams or N-word phrases of the form

\[
f_N(j) = \frac{F}{j^{1+\beta}} \ln^{N-1} \frac{\theta^N}{j^{1+\beta}}
\]

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 \( \mu_N \) and type/token-taken averages \( \mu^*_N \) 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 linguistic 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.
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

\[ P_r = \frac{B}{r^\beta} \]  

where \( r \geq 1 \). 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
\[ f_n(j) = \frac{F}{j^{1+\frac{1}{\beta}}} \ln^{N-1} \left( \frac{\sum G \theta}{j \theta} \right) \]  

(2)

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

\[ f_1(j) = \frac{F}{j^{1+\frac{1}{\beta}}} = \frac{F}{j^\alpha} \]  

(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 \( \mu_n \) and average use \( \mu_n^* \) 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 \( f : \{1, \ldots, \rho_m\} \rightarrow \mathbb{R} \) where \( f(j) \) denotes the density of sources with item density \( j \) and where \( \rho_m \) 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 \( g : \mathbb{R} \rightarrow \mathbb{R} \), where \( g(r) = \int_{\rho_m}^{\infty} f(j) \, dj \) 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 \( g \)):

\[
g^{-1}(j) = r(j) = \int_{\rho_m}^{\infty} f(j') \, dj'
\]

and also

\[
f(j) = -\frac{1}{g(g^{-1}(j))}.
\]

We also have that

\[
\int_{0}^{T} g(r) \, dr = A,
\]

the total number of items, hence

\[
P(r) = \frac{g(r)}{A}
\]

is the actual rank-frequency distribution.
In this framework, the law of Zipf is given by (since $r \in [0, T]$)

$$g(r)= \frac{E}{(1+ r)\beta}$$  \hspace{1cm} (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):

$$P(r) = \frac{D}{r^\beta}$$  \hspace{1cm} (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

$$T = \int_r f(j) \text{dj}$$  \hspace{1cm} (10)

and it can easily been proved that

$$A = \int_r f(j) \text{dj}.$$  \hspace{1cm} (11)

Hence

$$\mu = \frac{A}{T} = \frac{\int_r f(j) \text{dj}}{\int_r f(j) \text{dj}}$$  \hspace{1cm} (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 $\mu$ 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

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

for all $j \in [1, \rho_m]$ (note that the item densities – in practice 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 $\mu^*$, is now given by

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

where $A$ is as in (11) and where

$$W = \hat{O}_1^{\rho_m} jf(j) dj \quad (15)$$

In Egghe (2003), it is generally proved that

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

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 \( (\mu^*) \), encounter more books agreeing with a certain N-gram, than what could be expected from the average \( (\mu) \) 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 \( f, g, P, A, T, \mu, \mu^* \) 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

\[
P(r_i | r_{i+1}, \ldots, r_N) = P_i(r_i)
\]

\[17\]

i.e. the probability to have a letter or a word with rank \( r_i \) on the \( i^{th} \) 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 \( P_i(r_i) \)) and where we put \( P(r_1) = P_1(r_1) \).
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 \( P_i(r) \) for \( P(r) \)). Denote by \( P_N(r) \) the rank-frequency probability density function of N-word phrases or N-grams. Then

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

where \( \xi_N^{-1} \) denotes the inverse of \( \xi_N \) and where \( \xi_N \) is the function

\[
\xi_N(y) = \frac{\sum_{i=0}^{N-1} (-1)^{N+i-1} y \ln^i y}{\hat{a}}
\]

and \( \ln^i y = \ln(y) \cdot \ln(y) \cdot \ln(y) \ldots \) the \( i \)th power of \( \ln(y) \).

**Proof:**

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

\[
r = \text{vol} \left\{ (r_1,\ldots,r_N) | P(r_1,\ldots,r_N)^3 \times x \right\}
\]

where \( P(r_1,\ldots,r_N) \) denotes the probability of occurrence of an N-gram or N-word phrase for which the \( i \)th 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:

\[ P(r_1, \ldots, r_N) \]

\[ = P(r_1 \mid r_2, \ldots, r_N) P(r_2, \ldots, r_N) \]

\[ = P(r_1 \mid r_2, \ldots, r_N) P(r_2 \mid r_3, \ldots, r_N) P(r_3, \ldots, r_N) \]

\[ = \ldots \]

\[ = P(r_1 \mid r_2, \ldots, r_N) P(r_2 \mid r_3, \ldots, r_N) \ldots P(r_{N-1} \mid r_N) P(r_N) \]

\[ = P_1(r_1) P_1(r_2) \ldots P_1(r_{N-1}) P_1(r_N) \]

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

\[ r = \text{vol} \{ (r_1, \ldots, r_N) \mid P_1(r_1) P_1(r_2) \ldots P_1(r_N) \} \times x \]  

(21)

with \( x = P_N(r) \times \hat{0} [0, 1] \).

Note that, because of (8) and (9), the real ranks \( r_i \) should be lowered with 1 but, in (9), we can work with \( r_i \hat{0} [1, T+1] \) 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 \( r_i \) 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 \( r \hat{0} \frac{C}{C} \), then and \( r = T^N \) is obtained for \( x = 0 \), the set S being \( S = [1, T+1] \) which volume is \( T^N \) and \( r = 0 \) is obtained for \( x = \frac{\text{vol}(S)}{A \cdot C} \) since then \( \text{vol}(S) = 0 \) for the following reason: using (21) we have
\[ P_1(r_1)P_2(r_2)...P_N(r_N)^3 \times \frac{\alpha E \delta^N}{\beta A} \]

But by (7) and (8), each \( P_i(r_i) \leq \frac{E}{A} \). So

\[ P_1(r_1)P_2(r_2)...P_N(r_N) = \frac{\alpha E \delta^N}{\beta A} \]

But \( 0 \leq P_i(r_i) \leq \frac{E}{A} \) for every \( i = 1,...,N \) hence

\[ P_i(r_i) = \frac{E}{A} \quad (22) \]

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

\[ r_i = 1 \quad (23) \]

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

The inequality

\[ P_1(r_1)P_2(r_2)...P_N(r_N)^3 \times \]

leads to, using (9)

\[ \frac{D^N}{(r_1 r_2 r_3)^{\beta}} \times \]

hence
$r_1 r_2 \ldots r_N \leq \frac{N^\beta}{D_i^{\beta}} = \alpha$ \hspace{1cm} (25)

, by notation of $\alpha$ for reasons of simplicity. Formula (25) implies

$$1 \leq r_1 \leq \frac{\alpha}{r_2 \ldots r_N}$$ \hspace{1cm} (26)

This gives us the number of possible $r_1$s but dependent on the different $r_2$s, $\ldots$, $r_N$s that are possible. This will be determined now. Formula (26) yields

$$1 \leq r_2 \leq \frac{\alpha}{r_3 \ldots r_N}$$ \hspace{1cm} (27)

Formula (27) implies

$$1 \leq r_3 \leq \frac{\alpha}{r_4 \ldots r_N}$$ \hspace{1cm} (28)

and so on until

$$1 \leq r_{N-1} \leq \frac{\alpha}{r_N}$$ \hspace{1cm} (29)

and

$$1 \leq r_N \leq \alpha$$ \hspace{1cm} (30)

So $\text{vol}(S)$ of (21) is found when we remark that $r_1$ ranges in an interval of length $\frac{\alpha}{r_2 \ldots r_N} - 1$ (by (26)), where each $r_2, \ldots, r_N$ range as indicated in (27)-(30). Hence
The evaluation of (31) is tedious but easy.

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

\[
r = \frac{r_{N-1}^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_{N-1}^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
(31)
\]

\[
(> \ 0 \ by \ (27)), \ \text{we have that}
\]

\[
(1) = - a \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
(33)
\]

But

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

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

\[
(1) = - a \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
(34)
\]

But

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]

\[
\frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \ldots \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1} \frac{\tau_1^{\frac{a}{\tau_1}}}{\tau_1}
\]
Note that each time the sign switches. This leads to

\[(I) = \frac{r_N}{r_N} \frac{a}{r_N} \frac{a}{r_N} \frac{a}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \frac{d r_N}{r_N} \]

\[(I) = a \frac{\delta}{r_N} \frac{(-1)^{N+1}}{(N-2)!} \]

\[(I) = \frac{(-1)^N a \ln N}{(N-1)!} \]

\[(I) = \frac{a \ln N}{(N-1)!} \]

(35)

Since \( a = \frac{D^N}{x^N} \), using (25).

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

\[(II) = - \frac{r_N}{r_N} \frac{a}{r_N} \frac{a}{r_N} \frac{a}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \frac{d r_N}{r_N} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]

\[(II) = - \frac{\delta}{r_N} \frac{\delta}{r_N} \frac{\delta}{r_N} (-1)^{N-1} \ln a \frac{3 \sum_{N} \delta r_N}{a \delta a} \]
In general we have \((j + 3 = 2, \ldots, N)\)

\[
(\text{II}) = \frac{a^j}{r_{j+3}} \frac{1}{2!} \ln \frac{r_{j+3}}{a} - \frac{a}{r_{j+3}} \ln \frac{r_{j+3}}{a} + \frac{a}{r_{N}} \ln \frac{r_{N}}{a} - \frac{a}{r_{N}} \frac{\hat{u}}{\hat{u}}(\text{II})_{j+3}
\]

Hence

\[
(\text{II}) = \frac{a^j}{(i+1)!} \ln \frac{r_{i+1}}{a} - \frac{a}{r_{N}} \ln \frac{r_{N}}{a} + \frac{a}{r_{N}} \frac{\hat{u}}{\hat{u}}(\text{II})_{j+3}
\]

\[
(\text{II}) = (-1)^{N-1} \frac{a}{(i+1)!} \ln \frac{r_{i+1}}{a} - (-1)^{N-1} a \ln a + (-1)^N a
\]

\[
(\text{II}) = (-1)^{N-1} \frac{a}{i!} (-1)^{i+1} \ln \frac{r_{i+1}}{a} + (-1)^N a \ln a + (-1)^{N-1} a + (-1)^N
\]

\[
(\text{II}) = (-1)^{N-1} \frac{a}{i!} (-1)^{i+1} \ln \frac{r_{i+1}}{a} + (-1)^N a \ln a + (-1)^{N-1} a + (-1)^N
\]
\[(\Pi) = \sum_{i=0}^{N-2} \left( \frac{(-1)^{N+i-1}}{i!} a \ln a \right) + (-1)^N \] (36)

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

\[
r = \frac{\hat{\alpha} \ln^{N-1} a}{(N-1)!} + \sum_{i=0}^{N-2} \left( \frac{(-1)^{N+i-1}}{i!} a \ln a \right) + (-1)^N \]

(37)

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

\[
r + (-1)^{N-1} = \sum_{i=1}^{N} \frac{\xi_N(y)}{D(P_N(r))^i} \]

(38)

where

\[
\xi_N(y) = \sum_{i=0}^{N-1} \left( \frac{(-1)^{N+i-1}}{i!} y \ln^i y \right)
\]

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

\[
\xi_N(y) = \sum_{i=0}^{N-1} \left( \frac{(-1)^{N+i-1}}{i!} \ln^i y \right) + \sum_{i=1}^{N-1} \left( \frac{(-1)^{N+i-1}}{(i-1)!} \ln^i y \right)
\]

\[
\xi_N'(y) = \sum_{i=0}^{N-1} \left( \frac{(-1)^{N+i-1}}{i!} \ln^i y \right) + \sum_{i=0}^{N-2} \left( \frac{(-1)^{N+i-1}}{i!} \ln^i y \right)
\]
\[ \xi_{N}(y) = \frac{\ln^{N-1} y}{(N-1)!} > 0 \tag{39} \]

on \( y^\dagger \), \(+\infty \). So \( \xi_{N} \) is a strictly increasing function on \([1, +\infty)\) and hence an injection. But

\[ a = \frac{D^\beta}{\frac{1}{\beta} \ln^{\beta} 1} \]

by (25), hence we can take the inverse of \( \xi_{N} \) in (38) yielding

\[ P_{N}(r) = \frac{D^N}{\left(\xi_{N}^{-1}(r + (-1)^{N-1})\right)^\beta} \]

where \( \xi_{N}^{-1} \) denotes the inverse of the function \( \xi_{N} \).

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

\[ P_{N}(r) \approx \frac{D^N}{\left(\chi_{N}^{-1}\left((N-1)r\right)\right)^\beta} \tag{40} \]

where \( \chi_{N}^{-1} \) 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

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

In other words, in the proof of the above theorem we only calculate (I) for $r$ and put (II) $\leftarrow 0$. 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

$$g_N(r) = \frac{A_N D_N}{\left( \xi_N^{-1}(r + (-1)^{N-1}) \right)^{b}}$$

for $r \in \mathbb{R}$, using Theorem III.1.
In the proof of Theorem III.1 we showed that $\zeta_N$ strictly increases, hence the same is true for $\zeta_N^{-1}$, so $g_N$ strictly decreases, using (42). From (39) it follows that $\xi_N(y) > 0$ and $\xi_N^{-1}(y) > 0$ on $\mathbb{R}$. 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

$$f_N(j) = \frac{C}{j^\beta} \left[ \ln \frac{\lambda_m(N)}{j} \right]$$

(43)
for \( j \hat{1} \hat{\rho}_m(N) \hat{u} \), where \( \rho_m(N) \) is the maximal item density in the case if N-grams or N-word phrases, given by

\[
\rho_m(N) = A_N D^N
\]  

(44)

and where \( C \) is the constant

\[
C = \frac{\rho_m(N) \beta}{\beta^N (N-1)!}. 
\]  

(45)

**Proof:**

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

\[
f_N(j) = - \frac{1}{g_N(g_N^{-1}(j))} 
\]  

(46)

for \( j \hat{1} \hat{\rho}_m(N) \hat{u} \) 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

\[
g_N'(r)\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^\beta + g_N(r)\beta\left(\xi_N^{-1}\left(r + (-1)^{N-1}\right)\right)^{\beta-1} \frac{d(\xi_N^{-1})}{dx}(x = r + (-1)^{N-1}) = 0
\]

where

\[
\frac{d(\xi_N^{-1})}{dx}(x = r + (-1)^{N-1})
\]
means: the derivative of the function $\xi^{-1}_N$ in the point $r + (-1)^{N-1}$. So

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

But, by (39)

$$g_N'(r)\xi^{-1}_N(r + (-1)^{N-1}) = \frac{-\beta g_N(r)}{(N-1)! \ln^{N-1}\left(\xi^{-1}_N(r + (-1)^{N-1})\right)}.$$  \hfill (48)

Now we use (42), yielding

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

So

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

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

$$\xi^{-1}_N(r + (-1)^{N-1}) = \frac{\beta A_N D^N}{j \phi}.$$  \hfill (51)
in the point \( r = g_N^{-1}(j) \). So (51) in (50) yields

\[
g_N(g_N^{-1}(j)) = \frac{-\beta A_N D^N}{\beta j^{N+1} \ln^N \frac{A_N D^N}{j^\beta (N-1)!}}
\]  

which yields, by (46) the result

\[
f_N(j) = \frac{\left(\frac{A_N D^N}{\beta j^{N+1}}\right)^{1 \pm} \ln^N \frac{A_N D^N}{j^\beta}}{j^\beta (N-1)!}
\]  

\( j \hat{\xi}, \rho_m(N) \) a remarkably simple result. By definition of \( \rho_m(N) \) and \( g_N \) we have

\[
\rho_m(N) = g_N(0) = \frac{A_N D^N}{\left(\xi_N^{-1}(\cdot 1)^{N-1}\right)^{\beta}}
\]  

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

\[
\rho_m(N) = A_N D^N
\]

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

\[
f_N(j) = \frac{C}{\beta j^{N+1}} \ln^N \frac{\rho_m(N)^{\beta}}{j^\beta}
\]

with \( C \) as in (45), hence we have proved (43), for \( j \hat{\xi}, \rho_m(N) \).

\[\sim\]
Note that, in terms of Lotka’s $\alpha$, see (3), we have that (43) also reads as

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

(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 $\mathbb{R}$. $\rho_m(N)$ is $A_N D^N \frac{\beta}{\rho}$.

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 $C = \frac{\rho_m}{\beta}$) 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)

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

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

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 $\mu$ (here denoted as $\mu_N$) of items per source and $\mu^*$ (here denoted as $\mu_N^*$) 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 \( f_i \); in the next section we will use \( f_N \) \( (N^3 \cdot 2) \). Of course, the general defining formulae for \( \mu \) and \( \mu^* \) (i.e. for general size-frequency functions) of Section II also apply here.

V. Type/Token averages \( \mu_N \) and Type/Token-Taken averages \( \mu^*_N \) for N-grams and N-word phrases

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

\[
\mu_N = \frac{A_N}{T^N} \tag{56}
\]

\[
\mu^*_N = \frac{W_N}{A_N} \tag{57}
\]

where

\[
T^N = \int_1^{\rho_{\alpha}(N)} f_N(j) dj \tag{58}
\]

\[
A_N = \int_1^{\rho_{\alpha}(N)} j f_N(j) dj \tag{59}
\]

\[
W_N = \int_1^{\rho_{\alpha}(N)} j^2 f_N(j) dj \tag{60}
\]

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)): 
\[
\int x^n \ln^m x \, dx = \sum_{k=0}^{m} \binom{m}{k} (-1)^k (m+1)(m-1)...(m-k+1) \frac{\ln^{m-k} x}{(n+1)^{k+1}}
\]  

valid for all \( n \notin \{ -1 \} \) and \( m \notin \mathbb{Y} \).

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 \( \rho_m \) 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

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

so

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

Using (19) we have

\[
T^N + (-1)^N = \sum_{i=0}^{N-1} \binom{N+1}{i} (-1)^{N+1} \left( A_N D^N \right)^{\frac{1}{\beta}} \ln^i \left( \xi_N \left( A_N D^N \right)^{\frac{1}{\beta}} \right)
\]

hence, by (44)

\[
T^N = (-1)^N + \sum_{i=0}^{N-1} \binom{N+1}{i} (-1)^{N+1} \left( \rho_m (N) \right)^{\frac{1}{\beta}} \ln^i \left( \xi_N \left( \rho_m (N) \right)^{\frac{1}{\beta}} \right)
\]  

valid for all \( N \notin \mathbb{Y} \) and all \( \beta > 0 \).
We are left with the calculation of (59) and (60), using (43). We have

\[ A_N = \frac{\rho_m(N)}{\hat{\rho}_m(N)} \int_{\beta}^{\beta+1} \frac{\rho_m(N)}{j^\beta} dj. \]  

(63)

Since

\[ \int_{\beta}^{\beta+1} \frac{\rho_m(N)}{j^\beta} dj = -\rho_m(N), \]

we have that

\[ \hat{\rho}_m(N) = \rho_m(N) \int_{\beta}^{\beta+1} \frac{\rho_m(N)}{j^\beta} dj \]

\[ = \rho_m(N) \int_{\beta}^{\beta+1} \frac{\rho_m(N)}{j^\beta} dj. \]  

(64)

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

\[ \int_{\beta}^{\beta+1} \frac{\rho_m(N)}{j^\beta} dj = -1 \int_{\beta}^{\beta+1} \frac{\rho_m(N)}{j^\beta} dj = -1 \int_{\beta}^{\beta+1} \frac{\rho_m(N)}{j^\beta} dj. \]

(65)

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)
\[ A_N = \]
\[
\prod_{j=1}^{\beta} \frac{1}{\beta^N(N-1)!} \left( \sum_{k=1}^{N} \frac{(-1)^k (N-1)\ldots(N-k+1) \ln^k \rho_m(N)}{\beta^k} \right)
\]
\[
\prod_{j=1}^{\beta} \frac{1}{\beta^N(N-1)!} \left( \sum_{k=1}^{N} \frac{(-1)^k (N-1)\ldots(N-k+1) \ln^k \rho_m(N)}{\beta^k} \right)
\]
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

\[ A_N = \frac{\rho_m(N)}{\beta} \left( \frac{C}{j} \right) \ln^N \rho_m(N) \]

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

For \( W_N \) we have

\[ W_N = \frac{\rho_m(N)}{\beta} \left( \frac{C}{j} \right) \ln^N \rho_m(N) \]

But, using (63), we have

\[ \dot{\rho} \left( \frac{\rho_m(N)}{\beta} \right) \frac{1}{\beta} \ln^N \rho_m(N) \]

which can be calculated, using (61) for all \( \beta \neq \frac{1}{2} \). This gives
\[
\ln^{N-1} \frac{\mathcal{C}_m(N)}{j^{1/2}} \frac{\hat{a}(-1)^k N(N-1)\ldots(N-k)}{j^{1/2} N} \frac{\ln^{N-k} \frac{\mathcal{C}_m(N)}{j^{1/2}}} {j^{1/2} N} = \int_{j^{1/2}}^{j^{1/2} N} \frac{\hat{a}(-1)^k (N-1)\ldots(N-k+1)}{j^{1/2} N} \ln^{N-k} \frac{\mathcal{C}_m(N)}{j^{1/2}} \, dj
\]

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

\[
W_N = \frac{\left(\rho_\beta(N)\right)^{1/2}}{\beta^N (N - 1)!} \frac{(-1)^N (N - 1)!}{\hat{a}} \frac{\mathcal{C}_m(N)}{j^{1/2}} \frac{\ln^{N-k} \frac{\mathcal{C}_m(N)}{j^{1/2}}} {j^{1/2} N} = \int_{j^{1/2}}^{j^{1/2} N} \frac{\hat{a}(-1)^k (N-1)\ldots(N-k+1)}{j^{1/2} N} \ln^{N-k} \frac{\mathcal{C}_m(N)}{j^{1/2}} \, dj
\]

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

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

\[
\int_{j^{1/2}}^{j^{1/2} N} \frac{\hat{a}(-1)^k (N-1)\ldots(N-k+1)}{j^{1/2} N} \ln^{N-k} \frac{\mathcal{C}_m(N)}{j^{1/2}} \, dj = - \frac{\ln^{N-k} \frac{\mathcal{C}_m(N)}{j^{1/2}}} {j^{1/2} N}
\]

So (67) and (45) yield
\[ W_N = \frac{2^N (\rho_m(N))^2}{N!} \ln^N (\rho_m(N)) \]  

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

With these formulae for \( T^N, A_N \) and \( W_N \) we are able to calculate \( \mu_N \) and \( \mu_N^* \) via (56) and (57). We will also compare these values with the corresponding values of \( \mu_1 \) and \( \mu_1^* \), 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)

\[ T^2 = 1 - \rho_m(2) + \rho_m(2)\ln(\rho_m(2)) \]  

\[ A_2 = \frac{1}{2} \rho_m(2)\ln^2(\rho_m(2)) \]  

\[ W_2 = (\rho_m(2))^2 - \rho_m(2)\ln(\rho_m(2)) - \rho_m(2) \]  

Hence

\[ \mu_2 = \frac{\rho_m(2)\ln^2(\rho_m(2))}{2(1 - \rho_m(2) + \rho_m(2)\ln(\rho_m(2)))} \]
\[
\mu_2^* = \frac{2(\rho_m(2) - \ln(\rho_m(2)) - 1)}{\ln^2(\rho_m(2))}
\]  

(75)

which yields Table 1.

Table 1. Values of \(\mu_2\) and \(\mu_2^*\) for diverse
values of \(\rho_m(2)\), for \(\beta = 1\)

<table>
<thead>
<tr>
<th>(\rho_m(2))</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu_2)</td>
<td>1.140</td>
<td>1.244</td>
<td>1.397</td>
<td>1.600</td>
<td>1.890</td>
<td>2.933</td>
</tr>
<tr>
<td>(\mu_2^*)</td>
<td>1.150</td>
<td>1.277</td>
<td>1.494</td>
<td>1.846</td>
<td>2.526</td>
<td>8.902</td>
</tr>
</tbody>
</table>

This can be compared with the values of \(\mu_1\) and \(\mu_1^*\), i.e. the non-composed case. For \(\beta = 1\)
(hence \(\alpha = 2\)) we use the formulae (cf. Egghe (2003))

\[
\mu = \mu_1 = \frac{\ln \rho_m}{1 - \frac{1}{\rho_m}}
\]  

(76)

and

\[
\mu^* = \mu_1^* = \frac{\rho_m - 1}{\ln \rho_m}
\]  

(77)

yielding Table 2.
Table 2. Values of $\mu_i$ and $\mu_i^*$ for diverse values of $\rho_m$, for $\beta = 1$

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_i$</td>
<td>1.216</td>
<td>1.386</td>
<td>1.648</td>
<td>2.012</td>
<td>2.558</td>
<td>4.652</td>
</tr>
<tr>
<td>$\mu_i^*$</td>
<td>1.233</td>
<td>1.443</td>
<td>1.820</td>
<td>2.485</td>
<td>3.909</td>
<td>21.498</td>
</tr>
</tbody>
</table>

We see that, for the same value of the input “seed” $\rho_m (2)$ or $\rho_m$, we have that the values $\mu_i$ and $\mu_i^*$ are larger than the values $\mu_2$ and $\mu_2^*$ respectively. We also see that $\mu_2^* - \mu_2 < \mu_i^* - \mu_i$ 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 $\mu_i^* > \mu_i$ 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)

$$T_3 = - 1 + \rho_m (3) - \rho_m (3)\ln (\rho_m (3)) + \frac{1}{2} \rho_m (3)\ln^2 (\rho_m (3))$$  \hspace{1cm} (78)

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

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

Hence

$$\mu_3 = \frac{\rho_m (3)\ln^3 (\rho_m (3))}{-6 + 6\rho_m (3) - 6\rho_m (3)\ln (\rho_m (3)) + 3\rho_m (3)\ln^2 (\rho_m (3))}$$  \hspace{1cm} (81)

$$\mu_i^* = \frac{6\rho_m (3) - 3\ln^2 (\rho_m (3)) - 6\ln (\rho_m (3)) - 6}{\ln^3 (\rho_m (3))}$$  \hspace{1cm} (82)
which yields Table 3.

Table 3. Values of $\mu_3$ and $\mu_3^*$ for diverse values of $\rho_m(3)$, for $\beta = 1$

<table>
<thead>
<tr>
<th>$\rho_m(3)$</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_3$</td>
<td>1.103</td>
<td>1.179</td>
<td>1.288</td>
<td>1.431</td>
<td>1.630</td>
<td>2.329</td>
</tr>
<tr>
<td>$\mu_3^*$</td>
<td>1.110</td>
<td>1.200</td>
<td>1.348</td>
<td>1.577</td>
<td>1.989</td>
<td>5.148</td>
</tr>
</tbody>
</table>

The same comments as for $\mu_2$, $\mu_2^*$, given above, can be given here for $\mu_3$, $\mu_3^*$. Note again that the values of $\mu_3$, $\mu_3^*$ are smaller than the values of $\mu_2$, $\mu_2^*$ 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 $\mu_2$ and $\mu_2^{*}$ for diverse values of $\rho_m(2)$, for $\beta = \frac{1}{2}$

<table>
<thead>
<tr>
<th>$\rho_m(2)$</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_2$</td>
<td>1.130</td>
<td>1.214</td>
<td>1.321</td>
<td>1.433</td>
<td>1.552</td>
<td>1.761</td>
</tr>
<tr>
<td>$\mu_2^{*}$</td>
<td>1.140</td>
<td>1.244</td>
<td>1.397</td>
<td>1.600</td>
<td>1.890</td>
<td>2.933</td>
</tr>
</tbody>
</table>

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

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

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

yielding Table 5.

Table 5. Values of $\mu_1$ and $\mu_1^{*}$ for diverse values of $\rho_m$, for $\beta = \frac{1}{2}$

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>1.200</td>
<td>1.333</td>
<td>1.500</td>
<td>1.667</td>
<td>1.818</td>
<td>1.980</td>
</tr>
<tr>
<td>$\mu_1^{*}$</td>
<td>1.216</td>
<td>1.386</td>
<td>1.648</td>
<td>2.012</td>
<td>2.558</td>
<td>4.652</td>
</tr>
</tbody>
</table>
For $N = 3$, $\beta = \frac{1}{2}$ we have now, using (62), (65) and (67)

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

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

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

Hence

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

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

yielding Table 6.

**Table 6. Values of $\mu_3$ and $\mu_3^*$ for diverse values of $\rho_m(3)$, for $\beta = \frac{1}{2}$**

<table>
<thead>
<tr>
<th>$\rho_m(3)$</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_3$</td>
<td>1.097</td>
<td>1.160</td>
<td>1.241</td>
<td>1.330</td>
<td>1.429</td>
<td>1.635</td>
</tr>
<tr>
<td>$\mu_3^*$</td>
<td>1.103</td>
<td>1.179</td>
<td>1.288</td>
<td>1.431</td>
<td>1.630</td>
<td>2.329</td>
</tr>
</tbody>
</table>

We see again that the same tendencies of the comparison of $\mu_1$, $\mu_1^*$, $\mu_2$, $\mu_2^*$, $\mu_3$, $\mu_3^*$ 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


