Jointly published by Kluwer Academic Publishers, Dordrecht and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 47, No. 2 (2000) 237–252

The distribution of N-grams

LEO EGGHE

LUC, Diepenbeek (Belgium)
UIA, Wilrijk (Belgium)

N-grams are generalized words consisting of N consecutive symbols, as they are used in a text. This paper determines the rank-frequency distribution for redundant N-grams. For entire texts this is known to be Zipf's law (i.e., an inverse power law). For N-grams, however, we show that the rank (r)-frequency distribution is

$$P_N(r) = \frac{C}{(\psi_N(r))^{\beta}},$$

where ψ_N is the inverse function of $f_N(x)=x \ln^{N-1}x$. Here we assume that the rank-frequency distribution of the symbols follows Zipf's law with exponent β.

Introduction

N-grams are generalized words consisting of N consecutive symbols. N-grams are a very important study topic as they have applications in indexing, information retrieval, error correction, text compression, language identification, subject classification and even speech recognition.

Consider a text (in principle any kind of text) and fix $N \in \mathbf{N}$. Usually N=3, 4, 5 or 6 but this is not important here. In this text, using a "window" technique, one determines a (long) list of occurring N-grams (see, e.g., *Robertson* and *Willett*, 1998). Each type of N-gram is a coordinate in a vector, representing the text under study and the number appearing in this coordinate is the number of times this N-gram appears in the text. In this way texts can be compared by using similarity measures between the vectors of any two of these texts. In this way texts can be clustered when they are similar and this information is important in indexing. Classical similarity measures are: the Dice index (*Robertson* and *Willett*, 1998), the Jaccard index (*Hamers* et al., 1989; *Rousseau*, 1998) or Salton's cosine formula (*Salton* and *Mc Gill*, 1987) but other similarity measures can be used.

0138–9130/2000/US $ 15.00

This procedure can also be used on key words, hence associating with them similar key words (by applying a threshold on the similarity measure) to be used in indexing. Indexing and information retrieval (IR) are dual concepts (*Egghe* and *Rousseau*, 1997) and hence N-grams have applications in IR: clustered documents (as described above) are also retrieved and queries can be expanded by determining (as described above) similar key words. This technique is a considerable extension of the classical truncation technique.

Error correction can be done if a certain word is not recognized in a system (e.g. a word does not belong to the dictionary) and hence the word is considered to be in error. If not it is added as such ; in the other case similar words (determined by N-grams-similarity) are proposed for correction.

Language identification and subject classification are also cluster-based techniques, based on the fact that languages or topics (in a language) are characterized by their N-grams distributions. Similarities can be calculated and clusters (or other multivariable statistical methods) can be determined, hence presenting "atlasses" of languages or of topics (or scientific disciplines). In the same way can N-grams be used in the understanding of natural language.

Text compression is obtained by attaching heavily used N-grams to unused bytes which are reserved for single symbols (since the set of single symbols (hence 1-grams) usually is not an integer power of 2).

Finally, N-grams of phonemes or of words (not of letters) can be used in speech recognition. We refer the reader to the review articles of *Cohen* (1995), *Damashek* (1995), *Robertson* and *Willett* (1998) and *Yannakoudakis* et al. (1990) and to the many references therein. Note that all these techniques are language or subject independent (contrary to more classical ones, e.g. truncation, indexing, ...). It is however noted in the above references that these techniques perform better in Asian languages, because of their specific structure (symbols are entire syllables and hence truncation – to mention only this example – works not so well in these languages).

So, this should convince the reader that the study of N-grams is very important. One element in such a study is their rank-frequency distribution, i.e., fixing N, determine a relationship between the number of times a certain N-gram appears and its rank r (ranked in decreasing order of appearance). Hence we are interested in the N-gram analogue of the law of Zipf which is known to be valid in ordinary texts (which can be considered as texts consisting of N-grams but with variable N, i.e., the words of different lengths). As is the case with the law of Zipf, such a distribution determines the compressive power of N-grams. Indeed, the more concentrated (see *Egghe* and

Rousseau, 1990a,b; 1991) N-grams are (i.e., the more unequal their distribution is) the more effective compression that can be obtained.

Zipf's law is well-known. It states that

$$x = P(r) = \frac{C}{r^{\beta}}, \tag{1}$$

where r is the rank of a word (decreasing order of occurrence), P(r) is the fraction describing how often this word occurs and β and C are constants. Otherwise stated

$$r = \frac{C^{\frac{1}{\beta}}}{x^{\frac{1}{\beta}}}, \tag{2}$$

also a (decreasing) power law.

Is such a law also valid if one only considers N-grams (N fixed)? In the next section we prove that for redundant N-grams (see the next section for definitions)

$$r = D\frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}}\ln^{N-1}\left(\frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}}\right) \tag{3}$$

hence the composition of a "classical" Zipf law and the function

$$f_N(y) = y \ln^{N-1} y . \tag{4}$$

In other words, the exact law of Zipf is not valid in the case of redundant N-grams (N fixed). In the proof we assume that the symbols' rank-frequency distribution is Zipfian, a fact that can be accepted in a first approximation (also this is discussed in the next section).

This result is similar to a result obtained in *Egghe* (1999) on the distribution of N-word phrases but the present proof is a simplification of the argument given there and, in addition, is more general so that it also comprises N-grams distributions.

We note that, in our arguments for fixed N, we need less assumptions as, e.g., in the arguments of *Li* (1992) or of *Mandelbrot*, explaining his law for words in texts, based on symbols (see *Mandelbrot*, 1977; *Egghe* and *Rousseau*, 1990a). There one assumes that all symbols have an equal chance to occur. This is certainly not the case (a trivial fact), e.g., illustrated by Dewey's table appearing in *Heaps* (1978, p. 200). Using that all symbols have an equal chance to occur also implies that, the longer a word is the lesser

it is used! This is certainly not the case: words as the, an, it, ... are certainly used more heavily than letters such as j, z or q! Our model uses the real probabilities of the symbols and this distribution has a parameter influence on the final result. In *Perline* (1996) one also presents an argument using real symbol probabilities (although completely different from ours) but this argument seems to be in error (*Troll* and *beim Graben*, 1998).

The rank-frequency distribution of N-grams

In his analysis of theoretical texts, in which words are formed by symbols and blanks, *Mandelbrot* (1977) assumes the following simplifications of real life:

1. Symbols (mostly letters) follow independently in words,
2. All words (i.e., N-grams, $N \in \mathbf{N}$) are used,
3. All symbols have an equal probability of occurrence.

None of these assumptions is correct and this is so in any text and in any language. Also, condition 3 has the consequence that, the longer a word is, the less it is used. This is false in any language. Take the example of English: words like "the", "an", "in", "of" and so on are used more than the one-letter words "x" or "q" or "z". Also to look at this article, it is clear that the word "distribution" is used more than the word "cat".

So the novelty of the argument given in this article is – first of all – allowing the symbols to have a realistic distribution of occurrence, i.e., rank-frequency distribution. We have inspected the following sources:

(i) The data of Dewey (as given in *Heaps*, 1978, p. 200) on the occurrence of the 26 letters in English. The graph is given in Fig. 1.

(ii) The data, as extracted from one of my reports of the LUC-library management committee (in Flemish). The graph is given in Fig. 2.

(iii) The data extracted from the Website of *Beckman* (see *Beckman*, 1999) on the occurrence of more than 13,000 Chinese symbols. The graph is given in Fig. 3a and b for the first 300 and first 100 characters, respectively.

(iv) The three texts analysed by *Yannakoudakis* et al. (1990). One text is a book in modern Greek (here we have 24 letters) ; the two others are text books in English (one on the history of the Mormon church and one on database management systems). The graph of the Greek text is given in Fig. 4. The others are very similar.

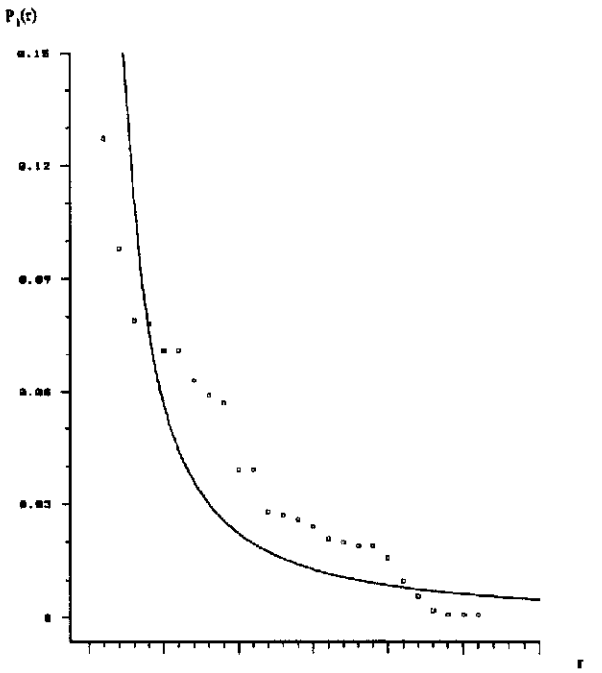

Fig. 1. Letter occurrence, Dewey data (English)

It is clear that, replacing the uniform distribution (as is the case in simplification (3)) by a decreasing power law of the type

$$P_1(r) = \frac{C}{r^{\beta}} \tag{5}$$

is a good first order assumption.

Note that (5) is equivalent with

$$r = \frac{C^{\frac{1}{\beta}}}{P_1(r)^{\frac{1}{\beta}}} \tag{6}$$

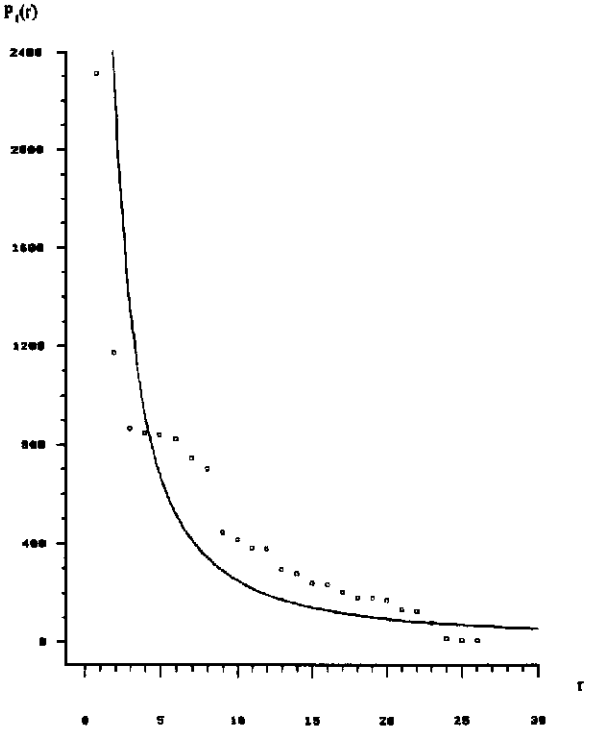

Fig. 2. Letter occurrence, LUC data (Flemish)

It is even almost perfect for the Chinese example (and we assume that this is also the case for the other Asian languages as well). Note that we remarked in the introduction that N-gram techniques work best in Asian languages.

This assumption also solves the simplification (2). Indeed, by allowing different probabilities for symbols in texts, we automatically add an extra diversification in the occurrence of words, exactly as they should appear. As to simplification (1) we note that, in general texts, it is certainly wrong but in the case studied here, it is true. This can be seen as follows. We will use N-grams generated by a window in an ordinary text.

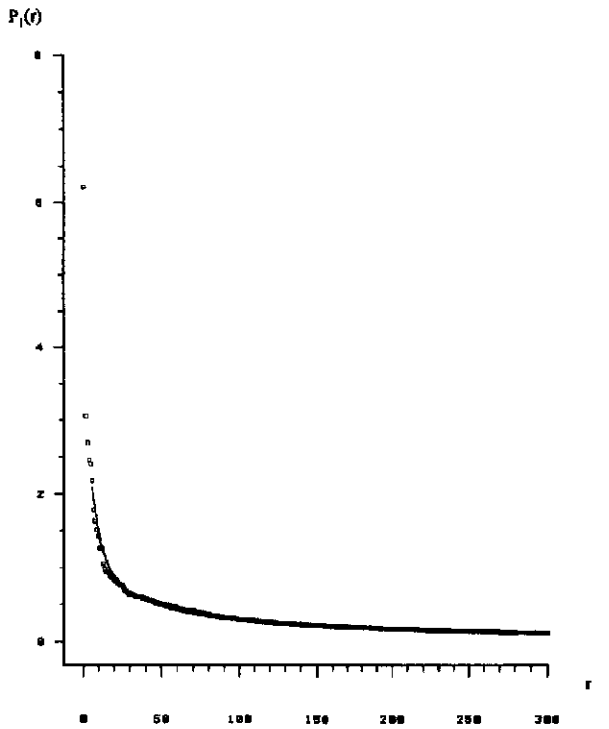

Fig. 3a. Occurrence of the 300 mostly used Chinese characters

We will use the redundant way of generation (see, e.g., *Willett*, 1979), the mostly used way to generate N-grams (see *Robertson* and *Willett*, 1998). To state the exact meaning of redundant generation of N-grams we show the generation from the word SYMBOL by 2- and 3-grams

redundant
2-Gram
S SY YM MB BO OL L
3-Gram
S *SY SYM YMB MBO BOL OL* L
non-redundant
2-Gram

SY MB OL
3-Gram
SYM BOL
(in the non-redundant case, incomplete N-grams are not used).

As is obvious from the redundant generation of N-grams, EVERY symbol (appearing in a word) appears once in the i^{th} place of an N-gram (i=1,...,N). Hence, the occurrence of the symbols in N-grams is independent of their place and of the symbols before them. In other words, the symbol distribution in any place i (i=1,...,N) in an N-gram is the same as the overall symbol distribution in the text. In mathematical notation: let r_i be the rank of the symbol in the i^{th} place in the N-gram (i=1,...,N).

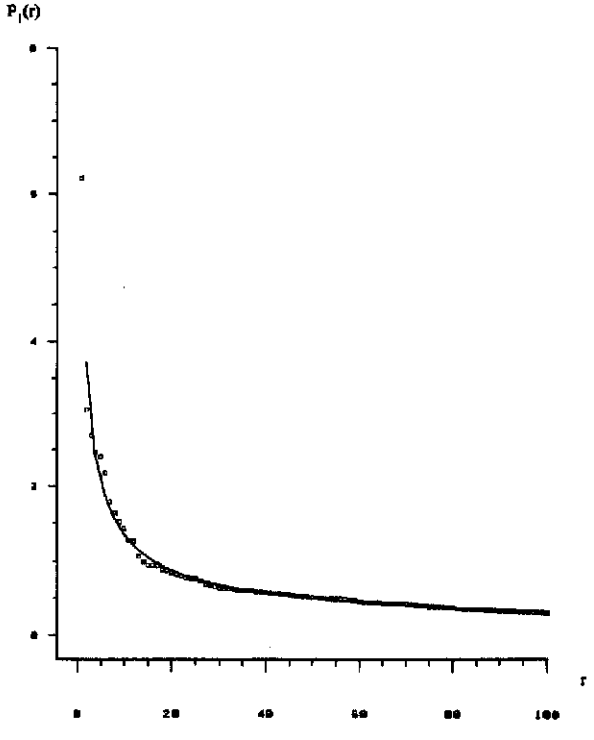

Fig. 3b. Occurrence of the 100 mostly used Chinese characters

Then, because of the above, the conditional probability to have r_i at the i^{th} place, given r_j at the j^{th} place (j=1,...,i-1) is

$$P(r_i|r_1,\ldots,r_{i-1}) = P_1(r_i), \tag{7}$$

where $P_1(.)$ denotes the symbol distribution in the text. For this distribution, we assume (5) (as explained above). The symbol set is assumed to be large enough so that also other symbols than letters can be used, or, in Asian languages, also all symbols can be considered. Since the alphabets in the latter cases are so high we can easily assume it to be infinite. In formula (5) we used the symbol P_1 denoting probabilities for symbols, i.e., 1-grams. Our task is, $N \in \mathbf{N}$ being fixed, to determine the distribution $P_N(r)$, denoting the probability for the N-gram on rank r to occur in the text. As ever, ranks are given according to the decreasing order of occurrence.

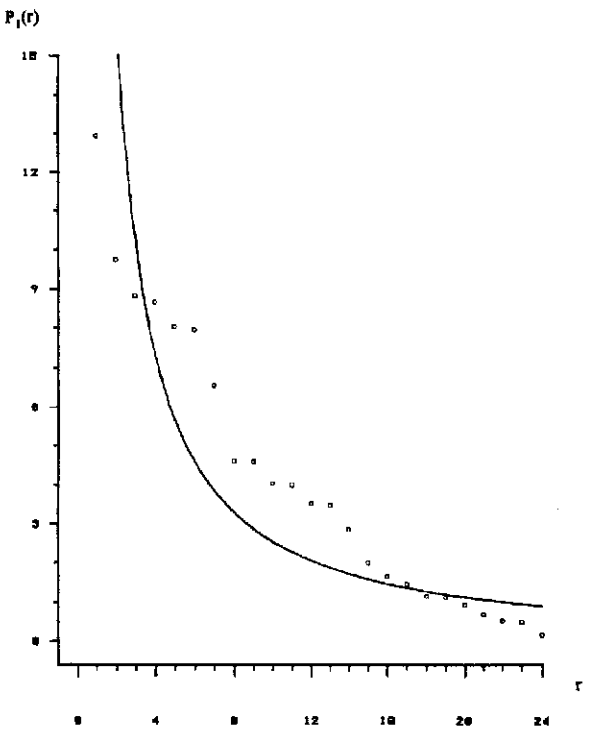

Fig. 4. Letter occurrence, ARGO1 data (Greek)

We have the following result.

Theorem 1. Let $N \in \mathbf{N}$ be fixed and assume (5) to be valid. Let $P_N(r)$ denote the probability (fraction) of occurrence of the N-gram on rank r. Then

$$r = \frac{C^{\frac{N}{\beta}}}{P_N(r)^{\frac{1}{\beta}}} \frac{\ln^{N-1}\left(C^{\frac{N}{\beta}} \Big/ P_N(r)^{\frac{1}{\beta}}\right)}{(N-1)!} \tag{8}$$

hence

$$r = \frac{1}{(N-1)!} f_N(y) , \tag{9}$$

where $f_N(y) = y \ln^{N-1}(y)$ and $y = \dfrac{C^{\frac{N}{\beta}}}{P_N(r)^{\frac{1}{\beta}}}$

(Note that for N=1, we refind (6)). Otherwise stated, there exist constants α_N and D_N such that

$$P_N(r) = \frac{D_N}{(\psi_N(\alpha_N r))^{\beta}} , \tag{10}$$

where $\psi_N = f_N^{-1}$. For large r this reduces to: there exists a constant E_N such that

$$P_N(r) = \frac{E_N}{(\psi_N(r))^{\beta}} , \tag{11}$$

hence a power law, but not of r but of $\psi_N(r)$. Note that the same exponent β (as in (5)) appears.

The proof is rather long, based on (5) and (7) and given in the Appendix. Nevertheless it is a simplification of the proof given in *Egghe* (1999) on the distribution of N-word phrases. The proof there can equally be simplified in this way. Note however that the result obtained in *Egghe* (1999) is an approximation of reality by assuming that the words in an N-word phrase occur independently from each other. The result obtained here for N-grams is exact since the letters indeed occur independently from each other (by (7)).

Conclusions and open problems

We have shown that, under more realistic assumptions than in *Mandelbrot* (1977), where we assume that the symbols follow a power law (instead of a uniform distribution) and where they occur independently of each other, that the N-gram distribution is a power law of $\psi_N(r)$, where ψ_N is the inverse of the function

$$f_N(x) = x \ln^{N-1} x .$$

This result is new and is obtained without any simplification (as mentioned in section II) or approximation.

As open problem we can formulate:

Establish a theory in which general texts (words with arbitrary lengths) can be treated.

*

The author is indebted to Prof. Dr. R. *Rousseau* for interesting discussions on this topic, to Mrs. V. *Kerstens* for the construction of some data sets and to Mrs. V. *Kerstens* and L. *Bruckers* for power law fitting of Chinese data.

References

BECKMAN (1999), http://casper.beckman.uiuc.edu/~c-tsai4/chinese/charfreq.html

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

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

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

EGGHE, L., ROUSSEAU, R. (1990a), *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science*, Elsevier, Amsterdam.

EGGHE, L., ROUSSEAU, R. (1990b), *Elements of concentration theory*, In: *Informetrics 89/90. Proceedings of the Second International Conference on Bibliometrics, Scientometrics and Informetrics, London (Canada)*, L. EGGHE, R. ROUSSEAU (Eds), Elsevier, Amsterdam, 97–137.

EGGHE, L., ROUSSEAU, R. (1991), Transfer principles and a classification of concentration measures, *Journal of the American Society for Information Science*, 42(7), 479–489.

EGGHE, L., ROUSSEAU, R. (1997), Duality in information retrieval and the hypergeometric distribution, *Journal of Documentation*, 53(5), 488–496.

HAMERS, L., HEMERICK, Y., HERWEYERS, G., JANSSEN, M., KETERS, H., ROUSSEAU, R., VANHOUTTE, A. (1989), Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula, *Information Processing and Management*, 25(3), 315–318.

HEAPS, H. S. (1978), *Information Retrieval, Computational and Theoretical Aspects*, Academic Press, New York.

LI, W. (1992), Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Transactions on Information Theory*, 38(6), 1842–1845.

MANDELBROT, B. B. (1977), *Fractals, Form, Change and Dimension*, W. H. Freeman and Company, San Francisco.

PERLINE, R. (1996), Zipf's law, the central limit theorem, and the random division of the unit interval, *Physical Review E*, 54(1), 220–223.

ROBERTSON, A. M., WILLETT, P. (1998), Applications of N-grams in textual information systems, *Journal of Documentation*, 54(1), 48–69.

ROUSSEAU, R. (1998), Jaccard similarity leads to the Marczewski-Steinhaus topology in IR, *Information Processing and Management*, 34(1), 87–94.

SALTON, G., MC GILL, M. J. (1987), *Introduction to Modern Information Retrieval*, Mc Graw-Hill, Singapore.

TROLL, G., P. BEIM GRABEN (1998), Zipf's law is not a consequence of the central limit theorem, *Physical Review E*, 57(2), 1347–1355.

WILLETT, P. (1979), Document retrieval experiments using indexing vocabularies of varying size II. Hashing, truncation, digram and trigram encoding of index terms, *Journal of Documentation*, 35(4), 296–305.

YANNAKOUDAKIS, E. J., TSOMOKOS, I., HUTTON, P. J. (1990), N-grams and their implication to natural language understanding, *Pattern Recognition*, 23(5), 509–528.

ZIPF, G. K. (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Harvard University Press, Cambridge (Massachussetts, USA).

Appendix

Proof of Theorem 1

It is clear from (7) that, for every $x \in]0,1[$, if we put

$$r = \#\{(r_1,...,r_N) \,\|\, P_1(r_1) \ldots P_1(r_N) \geq x\} \tag{A1}$$

we have that $x = P_N(r)$. Indeed: the probability for an N-gram for which the i^{th} symbol has rank r_i (i=1,...,N) is

$$P(r_N|r_1,...,r_{N-1})P(r_{N-1}|r_1,...,r_{N-2})...P(r_2|r_1)P(r_1) = P_1(r_N)P_1(r_{N-1})...P_1(r_1) \ ,$$

by (7). Note that in this paper N is fixed. We will consider $x \in]0,1[$ as a continuous variable, hence using continuous distributions $P_N(r)$, $r \geq 1$. In this sense (A1) should be interpreted as the volume of the N-dimensional set

$$\{(r_1,...,r_N) \,\|\, P_1(r_1) \ldots P_1(r_N) \geq x\} \ .$$

The inequality $P_1(r_1) \ldots P_1(r_N) \geq x$ leads to, using (5)

$$\frac{C^N}{(r_1 \ldots r_N)^\beta} \geq x \tag{A2}$$

hence

$$r_1 \ldots r_N \leq \frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}} \tag{A3}$$

To simplify this temporary situation denote

$$a = \frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}} \ . \tag{A4}$$

(A3) implies

$$1 \leq r_1 \leq \frac{a}{r_2 \ldots r_N} \ . \tag{A5}$$

This will give us the number of possible r_1s, but dependent on the number of different r_2s,...,r_Ns that are possible. This will be determined now. (A5) implies

$$1 \leq r_2 \leq \frac{a}{r_3 \dots r_N} . \tag{A6}$$

(A6) implies

$$1 \leq r_3 \leq \frac{a}{r_4 \dots r_N} \tag{A7}$$

and so on, until

$$1 \leq r_{N-1} \leq \frac{a}{r_N} \tag{A8}$$

and

$$r_N \leq a . \tag{A9}$$

We have by (A1), (A5)–(A9):

$$r = a \int_{r_N=1}^{r_N=a} \frac{dr_N}{r_N} \int_{r_{N-1}=1}^{r_{N-1}=\frac{a}{r_N}} \frac{dr_{N-1}}{r_{N-1}} \dots \int_{r_2=1}^{r_2=\frac{a}{r_3 \dots r_N}} \frac{dr_2}{r_2} \tag{A10}$$

After a long, inductive but easy calculation, (A10) resolves in:

$$r = \frac{a \ln^{N-1} a}{(N-1)!} \tag{A11}$$

By (A4) and the fact that $x = P_N(r)$, formula (7) follows. Hence also (8). (A11) now implies

$$\frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}} \ln^{N-1}\left(\frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}}\right) = (N-1)!r .$$

Hence

$$\frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}} = \psi_N((N-1)!r) , \tag{A12}$$

where $\psi_N = f_N^{-1}$ and where $f_N(y) = y\ln^{N-1}(y)$.

So

$$x = P_N(r) = \frac{D_N}{(\psi_N(\alpha_N r))^{\beta}}, \qquad (A13)$$

for $D_N = C^N$ and $\alpha_N = (N-1)!$, yielding (9). For large r, we proceed as follows from (A11):

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

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

But, for r high (equivalently x small), a is high by (A4). Hence

$$r^{\frac{1}{N-1}} \approx \left(\frac{a}{(N-1)!}\right)^{\frac{1}{N-1}} \ln\left(\frac{a}{(N-1)!}\right).$$

So

$$r \approx \frac{a}{(N-1)!} \ln^{N-1}\left(\frac{a}{(N-1)!}\right).$$

Hence

$$\frac{a}{(N-1)!} \approx \psi_N(r),$$

by definition of r. Consequently (10) follows for

$$E_N = \frac{C^N}{((N-1)!)^{\beta}},$$

using again that $a = \frac{C^{\frac{N}{\beta}}}{x^{\frac{1}{\beta}}}$. □

Received October 27, 1999.

Address for correspondence:
LEO EGGHE
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
UIA, Wilrijk, Belgium
E-mail : leo.egghe@luc.ac.be