General study of the distribution of N-tuples of letters or words based on the distributions of the single letters or words

GENERAL STUDY OF THE DISTRIBUTION OF N-TUPLES OF LETTERS OR WORDS BASED ON THE DISTRIBUTIONS OF THE SINGLE LETTERS OR WORDS

by

L. EGGHE

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

ABSTRACT

This paper establishes the general relation between the distribution of N-tuples of letters (e.g. N-truncations, N-grams) or words (e.g. N-word phrases) and the distributions of the single letters or words. Here the very general case is treated : the case where there is dependence on the place i in the N-tuple (i=1,...,N) in the sense that, for each i=1,...,N, a different distribution of the letters or words is supposed.
Concrete calculations are performed in the important case of Zipfian distributions (i.e. power laws) for the single letters or words. In this case we prove that the distribution of the N-tuples (N-fixed) is the sum of power laws.

[1] Permanent address.
Keywords and phrases : Zipf, N-word phrase, N-gram, N-truncation.

I. Introduction

The problem of studying the distribution of multi-word phrases (i.e. chains of words, say of fixed length $N \in \mathbb{N}$) has been introduced in [1] and solved in [2] in the case that, in an N-word phrase, the words appear independently of each other. Of course, as also noted in [2], this is not correct and the model only served as a first approximation. It must be noted that the model developed in [2] is already much more general than e.g. Mandelbrot's (see [3] or the more readable proof of it in [4] where one supposes (in the framework of N-tuples of letters) that the letters all have an equal chance to occur). This was not supposed in [2] and also not in [5] where one gives a proof of the distribution of N-grams (N fixed) based on the distributions of the letters. In this framework, independence of letters in N-grams is supposed but also proven to be exact in the case of redundant N-grams. For more details on N-grams (especially on their importance in virtually every aspect of information science) we refer the reader to [6], [7], [8] and [9] or to [5] in which these works are briefly described in the introduction.

The independence assumption may then well be true for N-grams, it is still an approximation of reality in the cases of N-word phrases and of texts consisting of words truncated to the length N (right of left truncated). Indeed, in the former case, words do not occur independently of each other and, in the latter case, in truncated words (as is the case in normal words in texts), the letters do not occur independently of each other. At the time of the writing of the papers [2], [5] it was not clear, however, how to deal with the dependent case and more in particular it was not clear to generalise the intricate proofs (as given in the appendices of [2] and [5]) to the dependent case.

This paper presents the solution to this problem, not in the least stimulated by a positive review of [2] by R. Marcus, see [10] in which the problem of knowing distributions of N-tuples of objects based on the objects' distributions has been recognized as very important. We present the solution in the next section (and the appendix).

Section III presents applications of this general result to the cases of N-tuples of letters and words : N-word phrases are discussed, texts of N-truncated words (i.e. left or right truncated words up to length N) are discussed as well as - for the sake of completeness -

redundant N-grams, where we repeat briefly the results of [5]. Section IV presents some conclusions and open problems.

II. The distribution of N-tuples of objects based on the distributions of the objects themselves - dependent case.

Since we want to capture the topics : N-word phrases and words consisting of N letters we will use henceforth the terminology : N-tuples of objects. In the case of N-word phrases the N objects are consecutive words ; in the case of words consisting of N letters, the objects clearly are consecutive letters. Note also that, in this paper, N will be fixed in $\mathbb{N}=\{1,2,3,...\}$.

For each $i\in\{1,...,N\}$ we denote by S_i the object set, i.e. the set of objects that can be used on the i-th place. Usually the S_is are equal (e.g. the alphabet, the set of possible words,...) but this is not supposed here. It is even so that the cardinals $\#S_i$ (i.e. the number of elements in S_i) do not have to be equal.

Denote, for each $i\in\{1,...,N\}$, $P(r_i|i)$ the rank-frequency distribution on "coordinate" i of the objects in S_i. So we clearly work here in a framework of i-dependence. Of course one more generalization could be possible, namely working with $P(r_i|r_1,...,r_{i-1})$, the probability to have the symbol in S_i on rank r_i, given that on the j^{th} place in the N-tuple $(j=1,...,i-1)$ we have the symbol on rank r_j in S_j. We leave open this very general case of dependence and will work only with dependence on the place $i\in\{1,...,N\}$. In other words, we put

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

for all $r_1,...,r_i$, being at least an important generalization of the independent case [there we would put $P(r_i|i) = P(r_i)$ where all S_i are equal and $r_i\in\{1,...,\#S_i\}$]. Let $P(r_1,...,r_N)$ denote the probability to have an N-tuple where, for each $i=1,...,N$, we use the object on rank r_i. By the very definition of conditional probability, applied repeatedly, we have

$$
\begin{aligned}
P(r_1,\ldots,r_N) &= P(r_N|r_1,\ldots,r_{N-1})P(r_1,\ldots,r_{N-1}) \\
&= P(r_N|r_1,\ldots,r_{N-1})P(r_{N-1}|r_1,\ldots,r_{N-2}).P(r_1,\ldots,r_{N-2}) \\
&= \ldots \\
&= P(r_N|r_1,\ldots,r_{N-1})P(r_{N-1}|r_1,\ldots,r_{N-2})\ldots P(r_2|r_1)P(r_1) \\
&= P(r_N|N)P(r_{N-1}|N\text{-}1)\ldots P(r_2|2)P(r_1) \qquad (2)
\end{aligned}
$$

, using (1).

Denote by $P_N(r)$ the rank-frequency distribution of the N-tuples. It takes only a moment's reflection to see that P_N is obtained as

$$x = P_N(r), \qquad (3)$$

where

$$r = \#\{(r_1,\ldots,r_N) \| P(r_1)P(r_2|2)\ldots P(r_N|N) \geq x\} \qquad (4)$$

Formulae (3) and (4) are the basic parts in the solution of the problem dealt with in this paper. They provide the device to derive the N-tuple distribution from the objects distributions. Of course, in order to be able to produce "concrete" functional relationships for (3) we must put in "concrete" functional relationships for the conditional probabilities $P(r_i|i)$, $i=2,\ldots,N$ and for $P(r_1)$. For the case of N-word phrases, we assume the word distributions to follow Zipfian distributions. It is well-known that this is the most evident assumption. But also in the case of N-tuples of letters, Egghe (1999) shows that Zipfian distributions can be used : especially for Asiatic languages the fit is almost perfect, but for the other languages the model is certainly acceptable. We therefore put

$$P(r_i|i) = \frac{C_i}{r_i^{\beta_i}}, \qquad (5)$$

$i=1,\ldots,N$ as the distributions of the objects on "coordinate" i. Here $r_i \in \mathbb{R}^+$. Indeed we will adopt the continuous setting for the ease of calculation. On the same lines, (4) will be

calculated by replacing # by vol, the volume of the N-dimensional set appearing in formula (4). We can prove the following theorem.

Theorem II.1 : Let $N\in\mathbb{N}$ be fixed and let (5) denote the conditional distributions of the objects on coordinate $i\in\{1,\ldots,N\}$. Then $x=P_N(r)$ is the distribution of the N-tuples, where

$$r = \sum_{j=1}^{N} \frac{D_j}{x^{1/\beta_j}} , \tag{6}$$

where the D_js are constants. Hence the inverse of (3) is a sum of Zipfian (i.e. power law) distributions.

Since the proof is elaborate and unrelated with the paper's topic we present it in the Appendix. Since theorem II.1 deals with the dependent case we supposed $\beta_i\neq\beta_j$ for all $i\neq j$. This will always be the case in practise. If some β_is are equal (but not all) we leave the result as an open problem. For the sake of completeness, if all β_is are equal, i.e. the Zipfian distribution of the objects is independent from the coordinate $i\in\{1,\ldots,N\}$, we repeat here the result of [5].

Theorem II.2 : Let $N\in\mathbb{N}$ be fixed and assume that

$$P(r_i) = \frac{C}{r_i^{\beta}} \tag{7}$$

denotes the object distribution (the same for all $i=1,\ldots,N$). Then $x=P_N(r)$ is the distribution of the N-tuples, where

$$r = \frac{D}{x^{1/\beta}} \frac{\ln^{N-1}(D/x^{1/\beta})}{(N-1)!} . \tag{8}$$

and where D is a constant. Hence

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

where $f_N(y) = y \ln^{N-1}(y)$ and $y = \frac{D}{x^{1/\beta}}$. It is shown in Egghe (1999) that (8) above, for large r, reduces to

$$P_N(r) \approx \frac{E}{(\psi_N(r))^{\beta}} \tag{9}$$

where E is a constant and where $\psi_N = f_N^{-1}$, the inverse of f_N.

Theorem II.1 above represents the case of dependence on the coordinate $i \in \{1,\ldots,N\}$ of the object distributions ; theorem II.2 represents the independent case. Both theorems have direct applications as will be seen in the next section.

III. Applications.

We start by repeating the independence result of Egghe (1999), where theorem II.2 can be used. We include it here for the sake of completeness.

III.1. The case of redundant N-grams

We refer to [5] and the already mentioned references on N-grams. Here we will suffice by explaining what redundant N-grams are. N-grams are, simply, words consisting of N letters. N-grams can be generated from normal texts by replacing each word (e.g. the word SYMBOL) by the string of N-grams, e.g. for N=2

⋆S SY YM MB BO OL L⋆

and e.g. for N=3

⋆⋆S ⋆SY SYM YMB MBO BOL OL⋆ L⋆⋆

The use of the stars makes sure that all letters in all words appear an equal number of times in the N-grams. This is the most important case for applications and are called redundant N-grams.

Note that in this case theorem II.2 applies since the distribution of the letters is the same (hence independent) on the coordinate $i \in \{1,\ldots,N\}$. We can conclude that redundant N-grams (as e.g. derived from texts) have a rank-frequency distribution as in (8), hence not a power law.

Note : A general note is in order here. Both theorems II.1 and II.2 show that the rank-frequency distribution P_N of N-tuples is not a power law, even when we supposed the object distributions to be power laws. In theorem II.1 we obtained for P_N essentially a power law of the inverse of the function $f_N(y)=y\ \ln^{N-1}(y)$. But since this last function ressembles very much a power law (in the sense of statistical fitting) it will be perfectly possible to fit $P_N(r)$ by a power law. The same remark goes for the result in theorem II.2 : there one essentially has that P_N is the inverse of a sum of power laws, which ressembles a power law in the statistical sense. In this paper we showed that - as a mathematical model - the power law is not correct for N-tuples, when the object distributions are power laws, in all cases (dependence on the place i or not).

III.2 The case of N-word phrases.

Here theorem II.1 applies since we can assume that words in N-word phrases occur according to Zipfian laws but where these laws are all different since words do not occur independently from each other. Hence (6) represents the rank frequency distribution (where $x=P_N(r)$) for N-word phrases.

This extends the result of [2].

III.3 The case of N-truncated words.

For e.g. information retrieval and indexing purposes it is interesting to know what is the distribution of truncated words (say up to length N, N fixed). Indeed the knowledge of such a distribution gives information on the inequality between occurrences of different N-truncated words. This, in turn, can be used in measures of concentration or diversity (as e.g. the entropy measure - see [11] or [12]) which are used (by applying some threshold on the concentration or diversity values) to determine the "selectivity" power of a key word.

Of course, contrary to the case of redundant N-grams, the occurence of letters in an N-truncated word is dependent on the place (the coordinate) in the N-tuple. Hence here theorem II.2 applies and formula (6) can be used. This is also the case for non-redundant N-grams.

IV. Conclusions, open problems

In this paper we have established the rank-frequency distribution of N-tuples of objects, given the conditional object distributions in coordinate $i \in \{1,\ldots,N\}$. We proved that this distribution is of the form $x = P_N(r)$, where

$$r = \sum_{i=1}^{N} \frac{D_i}{x^{1/\beta_i}} \tag{10}$$

in case the object distributions on coordinate i are Zipfian with exponent β_i, and where $i \neq j \Rightarrow \beta_i \neq \beta_j$. The case where all β_is are equal, dealt with in Egghe (1999), was repeated here for the sake of completeness.

We formulate here the problem of extending the conditional distributions $P(r_i|i)$ (as used here) to distributions of the form $P(r_i|i,N)$ where the probabilities also depend on the length N of the N-tuple and to prove results similar to the ones here but using $P(r_i|i,N)$ instead of the simpler $P(r_i|i)$. In this connection also a theory of truncation (right, left) would be interesting. A further extension could involve $P(r_i|r_1,\ldots,r_{i-1})$ for all $i=1,\ldots,N$.

It would also be nice to develop an axiomatic theory of texts which consist of words and where words consist of letters, i.e. a kind of "composition" of the separate results as described in section III.

References

1. R. Rousseau, A fractal approach to word occurrences in texts : the Zipf-Mandelbrot law for multi-word phrases. Preprint (1997).
2. L. Egghe, On the law of Zipf-Mandelbrot for multi-word phrases, *Journal of the American Society for Information Science* **50** (3), 233-241 (1999).
3. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York (1977).
4. L. Egghe and R. Rousseau, *Introduction to Informetrics*, Elsevier Science Publishers, Amsterdam (1990).
5. L. Egghe, The distribution of N-grams. Preprint (1999).
6. J.D. Cohen, Highlights : language- and domain-independent automatic indexing terms for abstracting, *Journal of the American Society for Information Science* **46** (3), 162-174 (1995).
7. M. Damashek, Gauging similarity with N-grams : language-independent categorization of text, *Science* **267** (10 February 1995), 843-848 (1995).
8. A.M. Robertson and P. Willet, Applications of N-grams in textual information systems, *Journal of Documentation* **54** (1), 48-69 (1998).
9. E.J. Yannakoudakis, I. Tsomokos and P.J. Hutton, N-grams and their application to natural language understanding, *Pattern Recognition* **23** (5), 509-528 (1990).
10. R. Marcus, *ACM Computing Review* **#339** (99-6-23) (1999).
11. H.S. Heaps, *Information Retrieval. Computational and Theoretical Aspects*, Academic Press, New York (1978).
12. G. Salton and M.J. Mc Gill, *Introduction to Modern Information Retrieval*, Mc Graw-Hill, Singapore (1987).

Appendix

Theorem : *Let* $N \in \mathbb{N}$ *be fixed and denote by*

$$P(r_i|i) = \frac{C_i}{r_i^{\beta_i}} \tag{A1}$$

the conditional distributions of the objects on coordinate $i \in \{1,\ldots,N\}$. Then $x = P_N(r)$ is the distribution of the N-tuples where

$$r = \sum_{j=1}^{N} \frac{D_j}{x^{1/\beta_j}} \quad , \tag{A2}$$

where the D_j s are constants.

Proof : In the continuous setting we have that $x = P_N(r)$ (the distribution of the N-tuples) iff (cf. (4))

$$r = \mathrm{vol}\{(r_1,\ldots,r_N) \| P(r_1)P(r_2|2)\ldots P(r_N|N) \geq x\} \tag{A3}$$

The inequality in the set in (A3) yields

$$\frac{C_1 \ldots C_N}{r_1^{\beta_1} \ldots r_N^{\beta_N}} \geq x \tag{A4}$$

Hence, denoting $a = \dfrac{C_1 \ldots C_N}{x}$

$$1 \leq r_1 \leq \frac{a^{1/\beta_1}}{(r_2^{\beta_2} \ldots r_N^{\beta_N})^{1/\beta_1}} \, . \tag{A5}$$

Hence the maximal value of r_1 is

$$\frac{a^{1/\beta_1}}{(r_2^{\beta_2} \ldots r_N^{\beta_N})^{1/\beta_1}} \quad ,$$

where $r_2,\ldots,r_N$ can vary as follows. (A5) implies

$$r_2^{\beta_2/\beta_1} \le \frac{a^{1/\beta_1}}{(r_3^{\beta_3}\ldots r_N^{\beta_N})^{1/\beta_1}}$$

Hence

$$1 \le r_2 \le \frac{a^{1/\beta_2}}{(r_3^{\beta_3}\ldots r_N^{\beta_N})^{1/\beta_2}} \tag{A6}$$

In the same way we have

$$1 \le r_3 \le \frac{a^{1/\beta_3}}{(r_4^{\beta_4}\ldots r_N^{\beta_N})^{1/\beta_3}} \tag{A7}$$

$$\vdots$$

$$1 \le r_{N-1} \le \frac{a^{1/\beta_{N-1}}}{(r_N^{\beta_N})^{1/\beta_{N-1}}} \tag{A8}$$

and

$$1 \le r_N \le a^{1/\beta_N} \tag{A9}$$

The above arguments yield

$$r = a^{1/\beta_1} \int_{r_N=1}^{r_N=a^{1/\beta_N}} \frac{dr_N}{r_N^{\beta_N/\beta_1}} \int_{r_{N-1}=1}^{r_{N-1}=\frac{a^{1/\beta_{N-1}}}{r_N^{\beta_N/\beta_{N-1}}}} \frac{dr_{N-1}}{r_{N-1}^{\beta_{N-1}/\beta_N}} \cdots \int_{r_2=1}^{r_2=\frac{a^{1/\beta_2}}{(r_3^{\beta_3}\ldots r_N^{\beta_N})^{1/\beta_2}}} \frac{dr_2}{r_2^{\beta_2/\beta_1}} \tag{A10}$$

We suppose here that $\beta_i \neq \beta_j$, $\forall i,j=1,\ldots,N$, $i \neq j$. We will evaluate the last integral, get an idea of what the general induction step will be and then complete the proof by complete induction.

$$\int_{r_2=1}^{r_2=\frac{a^{1/\beta_2}}{(r_3^{\beta_3}\ldots r_N^{\beta_N})^{1/\beta_2}}} \frac{dr_2}{r_2^{\beta_2/\beta_1}} = \frac{1}{1-\beta_2/\beta_1}\left[\frac{a^{1/\beta_2-1/\beta_1}}{r_3^{\beta_3(1/\beta_2-1/\beta_1)}\ldots r_N^{\beta_N(1/\beta_2-1/\beta_1)}} - 1\right]$$

Substituting this into (A10) yields the form

$$r = \gamma_2\, a^{1/\beta_2} \int_{r_N} \frac{dr_N}{r_N^{\beta_N/\beta_2}} \int_{r_{N-1}} \frac{dr_{N-1}}{r_{N-1}^{\beta_{N-1}/\beta_2}} \cdots \int_{r_3} \frac{dr_3}{r_3^{\beta_3/\beta_2}}$$

$$+ \gamma_1\, a^{1/\beta_1} \int_{r_N} \frac{dr_N}{r_N^{\beta_N/\beta_1}} \int_{r_{N-1}} \frac{dr_{N-1}}{r_{N-1}^{\beta_{N-1}/\beta_1}} \cdots \int_{r_3} \frac{dr_3}{r_3^{\beta_3/\beta_1}} \qquad \text{(A11)}$$

Here the bounds in all integrals are the same as in (A10) and

$$\gamma_2 = \frac{1}{1-\beta_2/\beta_1} = -\gamma_1$$

Suppose now, by complete induction that

$$r = \sum_{j=1}^{i} \gamma_j^{(i)}\, a^{1/\beta_j} \int_{r_N} \frac{dr_N}{r_N^{\beta_N/\beta_j}} \int_{r_{N-1}} \frac{dr_{N-1}}{r_{N-1}^{\beta_{N-1}/\beta_j}} \cdots \int_{r_{i+1}} \frac{dr_{i+1}}{r_{i+1}^{\beta_{i+1}/\beta_j}} \qquad \text{(A12)}$$

where $\gamma_j^{(i)}$ are constants. The last integral yields

$$\int_{r_{i+1}=1}^{r_{i+1}=\frac{a^{1/\beta_{i+1}}}{(r_{i+2}^{\beta_{i+2}}\ldots r_N^{\beta_N})^{1/\beta_{i+1}}}} \frac{dr_{i+1}}{r_{i+1}^{\beta_{i+1}/\beta_j}} = \frac{1}{1-\beta_{i+1}/\beta_j}\left[\frac{a^{1/\beta_{i+1}-1/\beta_j}}{r_{i+2}^{\beta_{i+2}/\beta_{i+1}-\beta_{i+2}/\beta_j}\ldots r_N^{\beta_N/\beta_{i+1}-\beta_N/\beta_j}} - 1\right] \qquad \text{(A13)}$$

Substituting this in (A12) and regrouping the terms yields (A12) but now for i replaced by i+1. Of course this also proves that

$$r = \sum_{j=1}^{N-1} \gamma_j^{(N-1)}\, a^{1/\beta_j} \int_{r_{N-1}}^{r_N=a^{1/\beta_N}} \frac{dr_N}{r_N^{\beta_N/\beta_j}} \qquad \text{(A14)}$$

which can be rewritten as (and denoting γ_j for $\gamma_j^{(N)}$)

$$r = \sum_{j=1}^{N} \gamma_j \, a^{1/\beta_j} \quad . \tag{A15}$$

Substituting the value of a and denoting the new constant D_j, this gives

$$r = \sum_{j=1}^{N} \frac{D_j}{x^{1/\beta_j}} \tag{A16}$$

completing the proof. Note that the β_j s are the Zipf-law powers with which we started in (A1). Law (A16), interpreted inversely as $x = P_N(r)$ is the rank frequency distribution for N-tuples. Note also that these calculations are exact in the sense that no approximations are used. □