

A formal derivation of Heaps' Law*

F.A. Grootjen and D.C. van Leijenhorst and Th. P. van der Weide

January 28, 2003

Abstract

In this paper Heaps' Law is derived from the Mandelbrot distribution.

1 Introduction

In many practical situations, a connection has been shown between the order of probability of events, and the probability itself. The most well-known models for such connections are Zipf's Law ([6]) and the Mandelbrot distribution ([4]). Let the r -th most probable event have probability p , then Zipf's law states that $p \cdot r$ is (almost) equal for all events, while the Mandelbrot distribution claims this for the expression $p \cdot (c + r)^\theta$ for some parameters c and θ . In case of $c = 0$, the distribution is also referred to as the generalized Zipf's Law. Some authors motivate the validity of these laws from physical phenomena, see for example [7] for Zipf's Law in the context of cities. But it is also possible to derive Zipf's law from a simple statistical model ([3]). For example, Zipf's Law can be derived for word occurrences in natural language, when it is assumed that words are drawn randomly from some distribution. In practice, however, words are thoughtfully selected by the author; yet on the long run this selection process may adjust to such a statistical description.

Another experimental law of nature is Heaps' Law ([2]), which describes the growth in the number of unique elements (also referred as the number of records), when elements are drawn from some distribution. Heaps' Law states that this number will grow according to αk^β for some application dependent constants α and β , where $0 < \beta < 1$. In the case of word occurrences in natural language, Heaps' Law predicts the vocabulary size from the size of a text.

In [1] Heaps' Law and the generalized Zipf's Law are related. It is shown that under a plausible assumption, it can be derived that if both Heaps' law and the generalized Zipf' Law hold, $\beta = 1/\theta$. The argumentation is as follows. Heaps' Law predicts the vocabulary size in a text of a given size. It is assumed that the frequency of the least frequent word in this text is $\Theta(1)$. As a result, the prediction of this frequency as obtained by the generalized Zipf's is also $\Theta(1)$. Working out the details leads to $\beta = 1/\theta$.

In this paper, we take another approach. We assume the generation of text is according to the Mandelbrot distribution, and derive Heaps' Law for the average vocabulary size in a text of a given length. This is done by a statistical analysis, leading to a rather untractable recurrence relation. As a consequence, Heaps' Law can also be regarded in a natural way as a complexity estimate. By applying techniques from complexity theory, restricting ourselves to first order terms, Heaps' Law is obtained. Note that by involving second order terms, a more advanced formulation of Heaps' Law may be obtained.

In figure 1 we see how nicely the function αk^β can be fitted against the function describing the average number records as a function of the number of drawings in the case of a set of 100 elements, while in figure 1 a set of 1000 elements is taken. This will be explained in more detail in section 2. It will be clear from these figures however that the approximation provided by Heaps'

*Technical Report NIII-R0302

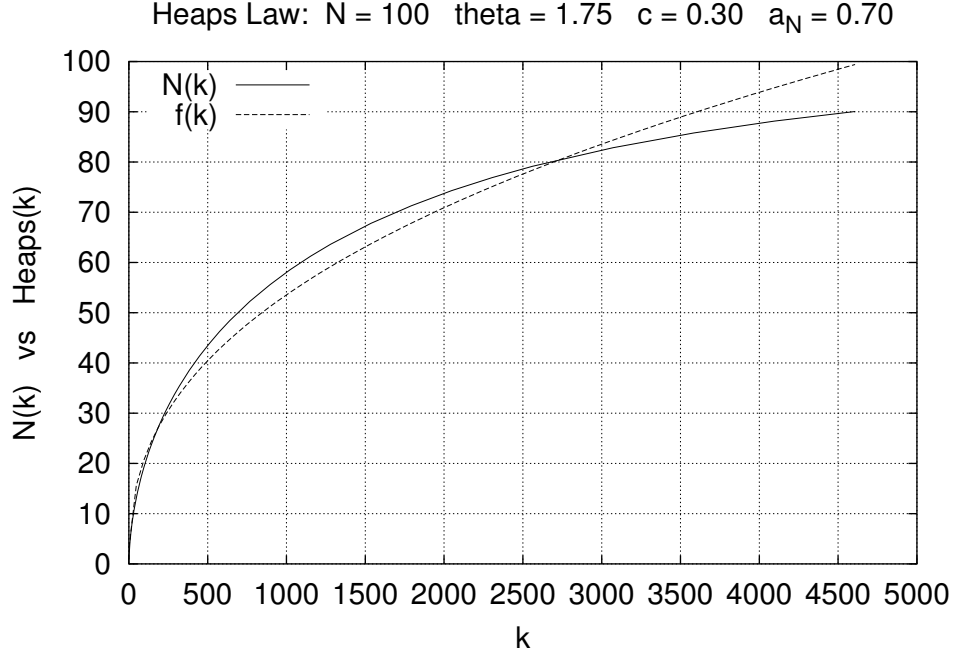


Figure 1: N_k versus Heaps law fit for $N = 100$

Law is not valid everywhere, as obviously the number of records is bounded by the total number of events, while a power function will exceed this number eventually.

In order to express this limited validity of Heaps' Law, we also focus on the validity area of the approximations in our analysis. The validity area is described rather defensively, in practice the area will be larger.

In section 2 we present a statistical model for the vocabulary size in a text, i.e. the average number of unique occurrences after a series of drawings. In section 3 we solve the resulting equation, leading to Heaps' Law. We also give bounds for the validity area of the approximations. In section 4 we make some conclusions and discuss further research.

2 A probabilistic model for Heaps' law

Let W be a set of N words numbered $1 \dots, N$, and let p_i the probability that word i is chosen. The underlying text model is that words are subsequently taken from the set W according to this probability distribution. We will be interested in the asymptotic behavior of the expected resulting number of different words taken.

After taking k words w_1, \dots, w_k from W , let $D_k = \{w_1, \dots, w_k\}$ be the set of different words, and let n_k be the number of such words: $n_k = \#D_k$. Then:

$$\begin{aligned}
 \text{Prob}(n_k = a) &= \text{Prob}(n_{k-1} = a - 1 \wedge w_k \notin D_{k-1}) + \text{Prob}(n_{k-1} = a \wedge w_k \in D_{k-1}) \\
 &= \text{Prob}(n_{k-1} = a - 1) * \text{Prob}(w_k \notin D_{k-1}) + \text{Prob}(n_{k-1} = a) * \text{Prob}(w_k \in D_{k-1})
 \end{aligned}$$

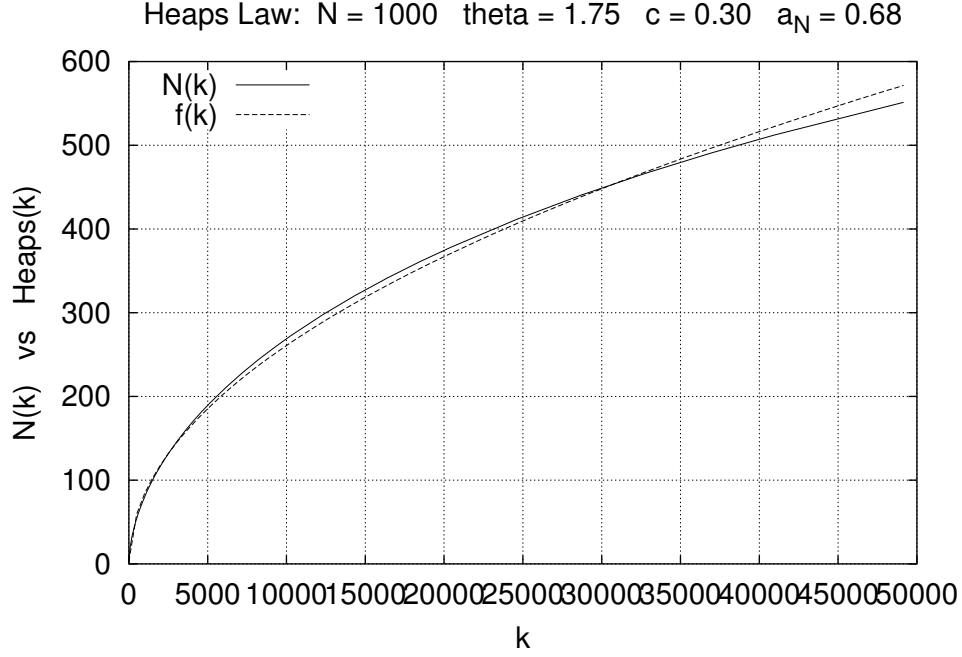


Figure 2: N_k versus Heaps' law fit for $N = 1000$

Then obviously:

$$\text{Prob}(w_k \notin D_{k-1}) = \sum_{i \in W} \text{Prob}(w_k = i \wedge i \notin D_{k-1}) \quad (1)$$

$$= \sum_{i \in W} p_i (1 - p_i)^{k-1} \quad (2)$$

Let $S_k = \sum_{i \in W} p_i (1 - p_i)^{k-1}$, and $M_n = \sum_{i \in W} (1 - p_i)^n$. We will refer to M_n as the k -th reverse moment of the probability distribution. Then we obtain $S_k = M_{k-1} - M_k$, as:

$$\sum_{i \in W} p_i (1 - p_i)^{k-1} = \sum_{i \in W} (-(1 - p_i)^k + (1 - p_i)^{k-1})$$

Let $N(k, a) = \text{Prob}(n_k = a)$, then we have $\sum_{a=1}^k N(k, a) = 1$. From the equations above we get the following recurrence relation::

$$\begin{aligned} N(1, 1) &= 1 \\ N(k, a) &= 0 \quad \text{if } k < a \\ N(k, a) &= N(k-1, a-1) * S_k + N(k-1, a) * (1 - S_k) \quad \text{if } k \geq a \end{aligned}$$

We will be interested in the expected number of different words N_k after taking k words randomly from the set W of words. Using this recurrence relation, we get for N_k ($k > 1$) the follow-

ing recurrence relation:

$$\begin{aligned}
N_k &= \sum_{a=1}^k a * N(k, a) \\
&= \sum_{a=1}^k a * (N(k-1, a-1) * S_k + N(k-1, a) * (1 - S_k)) \\
&= S_k \sum_{a=1}^k a * N(k-1, a-1) + (1 - S_k) \sum_{a=1}^k a * N(k-1, a) \\
&= S_k \left(\sum_{a=1}^k (a-1) * N(k-1, a-1) + \sum_{a=1}^k N(k-1, a-1) \right) + (1 - S_k) N_{k-1} \\
&= S_k * N_{k-1} + S_k * \sum_{a=1}^{k-1} N(k-1, a) + (1 - S_k) N_{k-1} \\
&= N_{k-1} + S_k
\end{aligned}$$

As a consequence:

Lemma 1 *The expected number of different words in a random selection of k words is $N_k = N - M_k$.*

Proof: $N_k = N_0 + \sum_{j=1}^k S_j = M_0 - M_k = N - M_k$ ◇

In order to estimate the asymptotic behavior of N_k , we will estimate M_k in the next section.

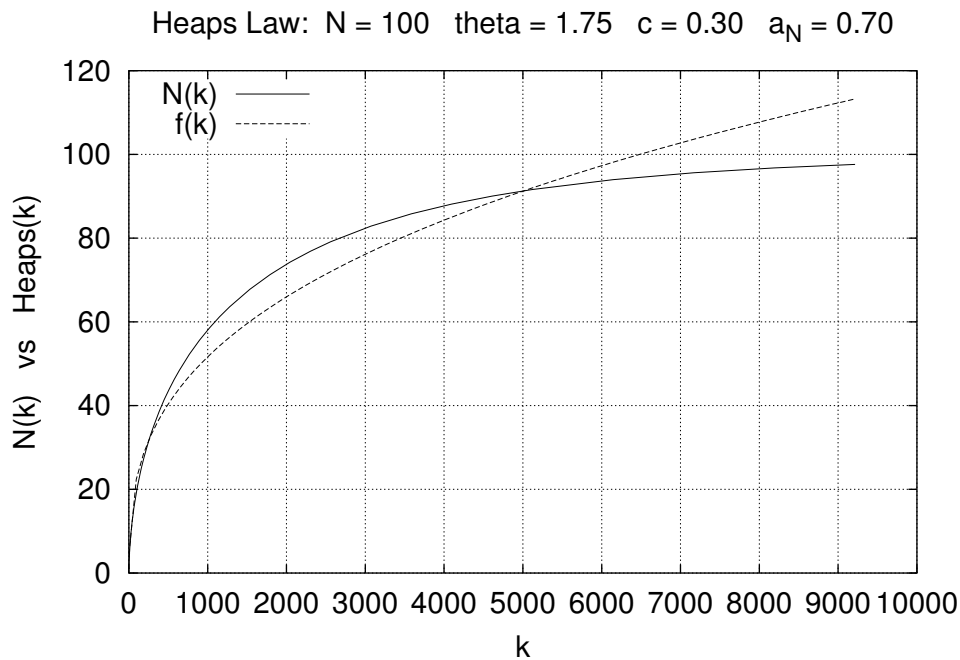


Figure 3: N_k versus Heaps law fit for $N = 100$

Applying curve fitting, we get some idea of the quality of this approximation provided by Heaps' Law. In figure 1 we see how Heaps' Law fits when 5000 elements are drawn from a set of 100 elements, while in figure 2 2000 drawings are taken. We see how the fit deteriorates. The same can be seen for larger values of N , for example, figure 1 shows the fits after 50000 drawings, and the worse situation in figure ?? after 500000 drawings.

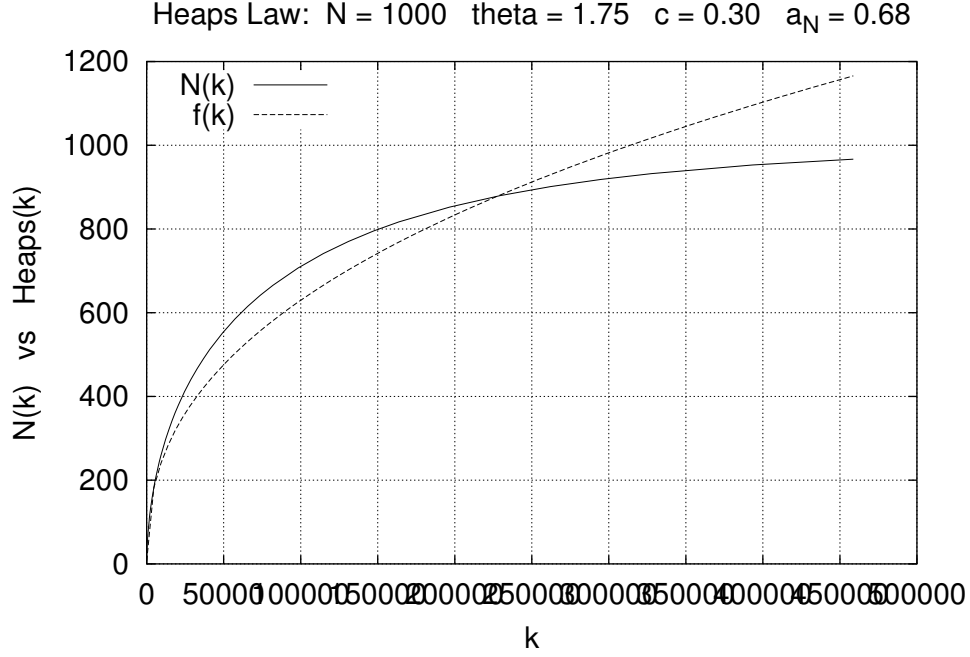


Figure 4: N_k versus Heaps law fit for $N = 1000$

3 Approximating reverse moments for Mandelbrot distribution

The Mandelbrot distribution provides a reasonable approximation of frequency of word usage in natural language. The Mandelbrot distribution assumes words to be ranked according to their frequency of usage. The probability of the word ranked at position i then corresponds to:

$$p_i = \frac{a_N}{(c + i)^\theta}$$

for some constants $c \geq 0$ and θ , where a_N is such that $\sum_{i=1}^N p_i = 1$. Usually the constant θ ranges over $[1, 2]$. As special case is Zipf's Law, which results from the Mandelbrot distribution by choosing $\theta = 1$ and $c = 0$. For the normalization constant we have:

Lemma 2 $\theta > 1 \implies a_N = \Theta(1)$

Let $t(x) = a_N(c + x)^{-\theta}$, then

Lemma 3 For $k > 0$ the function $\phi_k(x) = (1 - t(x))^k$ is increasing for $x \geq 1$.

- $\lim_{k \rightarrow \infty} \phi_k(x) = 0$
- $\lim_{x \rightarrow \infty} \phi_k(x) = 1$

The reverse moments may be estimated using the integral criterion:

Lemma 4

$$\sum_{i=1}^N \phi_k(i) = \int_1^N \phi_k(x) dx + \epsilon$$

with $\phi_k(1) \leq \epsilon \leq \phi_k(N) = (1 - a_N(c + N)^{-\theta})^k = \mathbf{o}(1)$.

The error ϵ of replacing summation by integration thus decreases exponentially in both k and N . Next we focus on estimating $\int_1^N \phi_k(x) dx$. By substituting $t = a_N(c+x)^{-\theta}$, leading to $dx = -At^\mu dt$ where $\mu = -1 - \beta$ and $A = \beta a_N^\beta$ and $\beta = \frac{1}{\theta}$, we get:

Lemma 5

$$\int_1^N \left(1 - \frac{a_N}{(c+x)^\theta}\right)^k dx = A \int_{t_N}^{t_1} (1-t)^k t^\mu dt$$

with $t_1 = a_N(c+1)^{-\theta}$, $t_N = a_N(c+N)^{-\theta}$.

We will process this outcome by applying partial integration:

Lemma 6

$$A \int_{t_N}^{t_1} (1-t)^k t^\mu dt = A \frac{(1-t)^k t^{\mu+1}}{\mu+1} \Big|_{t_N}^{t_1} - A\theta k \int_{t_N}^{t_1} (1-t)^{k-1} t^{\mu+1} dt$$

The first term of the righthand side approximates the number N of words in the set W :

Lemma 7

$$A \frac{(1-t)^k t^{\mu+1}}{\mu+1} \Big|_{t_N}^{t_1} = (c+N)\phi_k(N) - (c+1)\phi_k(1) = N + \Theta\left(\frac{k}{(c+N)^\theta}\right)$$

Proof:

$$\begin{aligned} A \frac{(1-t)^k t^{\mu+1}}{\mu+1} \Big|_{t_N}^{t_1} &= A \frac{(1-t_1)^k t_1^{\mu+1}}{\mu+1} - A \frac{(1-t_N)^k t_N^{\mu+1}}{\mu+1} \\ &= -(c+1)(1-t_1)^k + (c+N)(1-t_N)^k \\ &= (c+N)\phi_k(N) - (c+1)\phi_k(1) \end{aligned}$$

The result follows by the observation:

$$\phi_k(N) = \left(1 - \frac{a_N}{(c+N)^\theta}\right)^k = 1 + \Theta\left(\frac{k}{(c+N)^\theta}\right)$$

◇

For small values of k , the term $N\phi_k(N)$ is around N , but for larger values, this term decreases exponentially.

Next we focus on $A\theta k \int_{t_N}^{t_1} (1-t)^{k-1} t^{\mu+1} dt$.

Lemma 8

$$A\theta k \int_{t_N}^{t_1} (1-t)^{k-1} t^{\mu+1} dt = A\theta k \int_0^1 (1-t)^{k-1} t^{\mu+1} dt - \epsilon_2$$

where

$$\epsilon_2 = \Theta\left(\frac{k}{(c+N)^{\theta-1}}\right) + \mathcal{O}(k^2(1-t_1)^{k-1})$$

The latter term is exponentially decreasing in k .

Proof: Obviously

$$\epsilon_2 = A\theta k \int_0^{t_N} (1-t)^{k-1} t^{\mu+1} dt + A\theta k \int_{t_1}^1 (1-t)^{k-1} t^{\mu+1} dt$$

For the first term we notice that t_N will be almost 0 for large N and thus $1-t = \Theta(1)$. Consequently:

$$\begin{aligned} A\theta k \int_0^{t_N} (1-t)^{k-1} t^{\mu+1} dt &= \Theta \left(A\theta k \int_0^{t_N} t^{\mu+1} dt \right) \\ &= \Theta \left(a_N^\beta \frac{t_N^{1-\beta}}{1-\beta} k \right) \\ &= \Theta \left(\frac{a_N}{1-\beta} \cdot \frac{k}{(c+N)^{\theta-1}} \right) \end{aligned}$$

For the second term another application of partial integration is performed:

$$A\theta k \int_{t_1}^1 (1-t)^{k-1} t^{\mu+1} dt = A\theta k (1-t)^{k-1} \frac{t^{\mu+2}}{\mu+2} \Big|_{t_1}^1 + \frac{A\theta k(k-1)}{\mu+2} \int_{t_1}^1 (1-t)^{k-2} t^{\mu+2} dt$$

Both terms are exponentially decreasing in k . For the first part we have:

$$A\theta k (1-t)^{k-1} \frac{t^{\mu+2}}{\mu+2} \Big|_{t_1}^1 = -\frac{A\theta t_1^{\mu+2}}{\mu+2} \cdot k(1-t_1)^{k-1}$$

Note that $k(1-t_1)^{k-1}$ decreases exponentially to zero. This is also the case for the second part:

$$\begin{aligned} \frac{A\theta k(k-1)}{\mu+2} \int_{t_1}^1 (1-t)^{k-2} t^{\mu+2} dt &\leq \frac{A\theta k(k-1)}{\mu+2} \int_{t_1}^1 (1-t)^{k-2} dt \\ &\leq \frac{A\theta}{\mu+2} \cdot k(k-1)(1-t_1)^{k-1} \end{aligned}$$

◇

We proceed with $A\theta k \int_0^1 (1-t)^{k-1} t^{\mu+1} dt$, and recognize this integral as the Beta-function $B(\mu+2, k)$.

Lemma 9

$$A\theta k \int_0^1 (1-t)^{k-1} t^{\mu+1} dt = A\theta k B(k, \mu+2)$$

The Beta-function can be expressed in terms of the Gamma-function:

$$B(k, \mu+2) = \frac{\Gamma(k)\Gamma(\mu+2)}{\Gamma(k+\mu+2)}$$

Note that expression is only valid for $\mu+2 \neq 0$, which is equivalent with $\theta \neq 1$. In this paper we restrict ourselves to this case.

Substituting Stirling's approximation of the Γ -function $\Gamma(x+1) = \sqrt{2\pi x} x^x e^{-x} (1 + \mathbf{o}(1))$ (see [5]) on the terms $\Gamma(k)$ and $\Gamma(k + \mu + 2)$ yields:

$$\begin{aligned}
& A\theta k \frac{\Gamma(k)}{\Gamma(k + \mu + 2)} \Gamma(\mu + 2) \\
& \sim A\theta k \frac{\sqrt{2\pi}(k-1)^{k-1} e^{-k+1} (k-1)^{\frac{1}{2}}}{\sqrt{2\pi}(k + \mu + 1)^{k+\mu+1} e^{-k-\mu-1} (k + \mu + 1)^{\frac{1}{2}}} \Gamma(\mu + 2) \\
& = a_N^\beta \cdot \left(\frac{k-1}{k + \mu + 1} \right)^k \cdot \left(\frac{k-1}{k + \mu + 1} \right)^{\frac{1}{2}} \cdot e^{\mu+2} \cdot (k + \mu + 1)^{-\mu-1} \cdot \left(1 - \frac{1}{k}\right) \cdot \Gamma(\mu + 2) \\
& \sim a_N^\beta e^{\mu+2} \Gamma(\mu + 2) (k + \mu + 1)^{-\mu-1} \\
& \sim a_N^\beta e^{\mu+2} \Gamma(\mu + 2) k^{-\mu-1} \\
& = a_N^\beta e^{1-\beta} \Gamma(1 - \beta) \cdot k^\beta
\end{aligned}$$

Summarizing we have

Lemma 10

$$\sum_{i=1}^N \phi_k(i) = N - \alpha \cdot k^\beta \cdot (1 + \mathbf{o}(1)) + \Theta\left(\frac{k}{(c + N)^\theta}\right)$$

where $\alpha = e^{1-\beta} \Gamma(1 - \beta)$

This leads to the main result:

Lemma 11 *The expected number of different words in a random selection of k words is*

$$N_k = \alpha \cdot k^\beta \cdot (1 + \mathbf{o}(1)) + \Theta\left(\frac{k}{(c + N)^\theta}\right)$$

And thus we have shown:

Theorem 1 (Heaps' Law) $N_k = \alpha \cdot k^\beta$, with validity interval restricted to $k = \Theta((c + N)^\theta)$.

4 Conclusions and further research

In this paper we have derived the Heaps' Law from the Mandelbrot distribution, and provided a validity area for Heaps' Law. As a next step, a second order approximation may be employed, providing a sharper formulation for Heaps' Law, and a larger validity area. Furthermore, other distributions may be examined, leading to *Heaps' Criterion* as a sufficient condition for a distribution to imply Heaps' Law.

References

- [1] Ricardo A. Baeza-Yates and Gonzalo Navarro. Block addressing indices for approximate text retrieval. *Journal of the American Society of Information Science*, 51(1):69–82, 2000.
- [2] H. S. Heaps. Information retrieval: Computational and theoretical aspects. pages 206–208, 1978.
- [3] Wentian Li. Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6) 1842-1845, 1992.

- [4] B. Mandelbrot. The pareto-levy law and the distribution of income. *International Economic Review*, I, pages 79–106, 1960.
- [5] Schaum. Schaums Handbook of Formulas and Tables.
- [6] G. Zipf. *Human Behavior and the Principle of Last Effort*. 1949.
- [7] Xavier Gabaix Zipf. Zipf's law for cities: An explanation*.