Statistica Sinica 2(1992), 189-201

FROM ANIMAL TRAPPING TO TYPE-TOKEN

Min-Te Chao

Academia Sinica

Abstract: Let V_n be the number of word types in a text of n words. If the arrival of the ith word type is governed by a Poisson process with rate λ_i, we show that the growth rate of the series $\Sigma\lambda_i$ determines the asymptotic behavior of V_n. Specific conditions are given for $V_n \to \infty$, its rate of divergence, as well as its asymptotic rate of convergence to a distribution. The cases that $\lambda_n = n^{-p}$, $p > 1$ and $\lambda_n = \alpha^n$, $0 < \alpha < 1$, are fully discussed, and they agree with two well known classical species-area models. Finally, for certain finite vocabulary cases, a correction factor is introduced.

Key words and phrases: Linguistics, Poisson process, regular variation, slow variation, species-area.

1. Introduction

Let V_n be the number of different word types in a text that contains n words. In statistical linguistics, V_n is called a type-token relation (Herdan (1960)), and it is one of the more interesting problems that concerns both linguists and statisticians. In terms of stochastic abundance models, V_n can be interpreted as the number of different species that have been found up to time n, or within a geographical area n, see Engen (1978). The function V_n also has other names in the literature: it is called the rarefaction curve in ecology (Tipper (1979), Walton (1986)), although the appropriateness of rare function as a model for ecological processes is controversial (Lewin (1983)). For small n, the distribution of V_n is derived by Emigh (1983) for the multinomial case, and by Walton (1986) for the hypergeometrical case. In this paper, we are more concerned with the behavior of V_n for large values of n. Traditionally, V_n is either obtained via empirical studies (Yule (1944), Guiraud (1959), also see Gani (1985) for additional references) or through theoretically based models (Brainerd (1982), Gani (1985), Daley, Gani and Ratkowsky (1988), Sichel (1986)). Except for Sichel's work, which tries to fit a probability distribution to the observed data of V_n, almost all theoretical models considered so far are based on homogeneous Markov chains with suitable transition probabilities. Brainerd (1982) made an excellent attempt to link the type-token approach, which often uses stochastic processes as a modeling device, and the species-area problem, which is often based on a more static probability model. By choosing a suitable Markov transition probability, Brainerd succeeded in establishing two classical species-area models: the Gleason (1922) model, and the Arrhenius (1921) model.

A basic assumption for the stochastic processes approach is that the process $\{V_n, n \geq 1\}$ is Markov. On the other hand, the static species-area model assumes a multinomial distribution. These two assumptions are inconsistent (McNeil (1973)). We are more inclined to adopt the basic view of Herdan (1966, p.15): that the relative frequencies of symbols appear to be a common characteristic of linguistic forms; i.e., the multinomial framework. To fully utilize this structure we shall introduce a time variable t, go one step further and use the basic species-trapping framework of Fisher, Corbet and Williams (1943) in which each word type appears in accordance with a Poisson process with rate λ_i. In this paper we approach the type-token problem directly from Fisher's framework and, for the most part, discuss the behavior of V_n for large values of n.

Our major findings can be summarized by saying that the asymptotic behavior of V_n depends largely on the tail behavior of the series $\sum \lambda_n$. For arbitrary $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots)$, it is possible to roughly estimate $E(V_n)$ or $\mathrm{Var}(V_n)$. For certain specific $\boldsymbol{\lambda}$, it is possible to derive $E(V_n)$, $\mathrm{Var}(V_n)$ or the distribution of V_n as a function of n. Two important cases deserve special mention. If $\lambda_n = n^{-p}$, $p > 1$, we obtain, for some finite C_1 and C_2,

$$\begin{aligned} E(V_n) &= C_1 \cdot n^{1/p}(1 + o(1)), \\ \mathrm{Var}(V_n) &= C_2 \cdot n^{1/p}(1 + o(1)) \quad . \end{aligned} \tag{1.1}$$

and also conclude that V_n, properly normalized, is asymptotically normal. If $\lambda_n = \alpha^n$, $0 < \alpha < 1$, then

$$V_n = C_3 \cdot \log n + O_p(1), \tag{1.2}$$

for some constant C_3, and there is no need to normalize V_n to obtain an asymptotic distribution. Explicit values of C_1, C_2 and C_3 can be found. The normal case has been treated by McNeil (1973) in a different way. Relation (1.2) appears to be new.

These two examples enable us, to a large extent, to classify the type-token curves according to the rate of convergence of λ_n. First, there is no linear rate although it can be arbitrarily approached. If $\lambda_n \downarrow 0$ polynomially with degree p, then V_n is $O_p(n^{1/p})$, $p > 1$; furthermore, V_n is asymptotically normal. If $\lambda_n \downarrow 0$ geometrically or faster, then $V_n = O(\log n) + O_p(1)$; and the distribution of the $O_p(1)$ term can be explicitly found in the sense that we can find all its moments.

2. Fisher's Model

We shall, for ease of presentation, adopt Fisher's species-trapping description as our basic model. There are infinitely (but countably) many species, and the arrival of the ith one is governed by a Poisson process $X_j(t)$ with rate λ_j. The number of different species discovered in the time period $(0,t)$ is given by the simple formula

$$V(t) = \sum_{j=1}^{\infty} I_{[X_j(t)\geq 1]}.$$

These Poisson processes are assumed mutually independent. This framework is also adopted by McNeil (1973) to estimate the vocabulary of an author, and by Efron and Thisted (1976) to study how many words Shakespeare knew. The Poisson assumption is not essential for most of the results that follow. Since we shall be mainly interested in the asymptotic behavior of $V(t)$, particularly when t is large, we shall make the assumption that $\sum \lambda_i < \infty$ and $\lambda_i \geq 0$ for all i. Note that this is a sensible assumption since if $\sum \lambda_i = \infty$, we would be observing infinitely many arrivals in any finite interval $(0,t)$. We shall make a point in noting that, by conditioning the Poisson processes on their arrival epochs only, we obtain the classical multinomial framework. It is easy to see that if $\lambda_i > 0$ for all i, then $E[V(t)] \to \infty$. In fact, the condition that $\lambda_i > 0$ infinitely often is necessary and sufficient for $V(t) \to \infty$ a.s. The proof is omitted.

A basic issue underlying the type-token problem is whether we assume that the vocabulary size is finite. While the finite assumption is realistic, nevertheless the infinite vocabulary assumption does provide an easier mathematical framework. The first two theorems below imply that if we are to discover any non-trivial properties of $V(t)$ at all and at the same time keep the true parameter λ fixed, it is necessary to assume $\lambda_i > 0$ infinitely often. Some authors bypass this issue. For example, McNeil (1973) assumes a finite vocabulary initially, and later lets the vocabulary size tend to ∞. He has to assume a special form of λ_n (McNeil (1973), Eq.(4.2)). Efron and Thisted (1976) allow a possibly infinite vocabulary, but pay the price by finding only the lower bound for Shakespeare's vocabulary size.

To study the asymptotic behavior of $V(t)$, we shall be mostly interested in the rate at which $V(t) \to \infty$. This depends on the structure of the sequence $\lambda = (\lambda_1, \lambda_2, \ldots)$; we shall write $V(t,\lambda)$ to emphasis its dependence when needed. The next result is simple but useful; we shall omit the proof.

Theorem 1. *If $\lambda \leq \mu$ in the sense that $\lambda_i \leq \mu_i$ for all i, then $V(t,\lambda) \leq V(t,\mu)$ for all t.*

Strictly speaking, $V(t,\lambda)$ implicitly depends on a sequence of Poisson processes and the expression that $V(t,\lambda) \leq V(t,\mu)$ is unjustified. But for each sequence of Poisson processes $X_j(t)$ with rate λ_j, there exists a Poisson sequence $Y_j(t)$ with rate μ_j such that $X_j(t) \leq Y_j(t)$. It is in this sense that the above theorem is interpreted.

Corollary. *If $\lambda \leq \mu$, then $E[(V(t,\lambda)] \leq E[V(t,\mu)]$.*

For certain λ, it is possible to determine the rate at which $V(t,\lambda)$ diverges. Theorem 3 below provides a useful basis for comparison to determine the rate of $V(t,\mu)$.

3. The Expected Number of Types

In Section 2 we have given a necessary and sufficient condition for $V(t) \to \infty$ a.s. This condition is independent of the growth rate of $\sum \lambda_i$. In this section, we shall study the behavior of $E[V(t,\lambda)]$, and try to determine, at least for large values of t, the approximate form of $E[V(t,\lambda)]$. A salient feature of our approach is that it tries to link a particular convergent series $\sum \lambda_n$ with the form of the type-token relationship. We shall assume $\lambda_n \downarrow 0$. For the moment it is assumed $\lambda_n > 0$ for all n.

Let $\lambda = (\lambda_1, \lambda_2, \ldots)$, $\lambda_n > 0$ for all n. Let $h_n = \lambda_n \{\sum_{i=1}^{\infty} \lambda_{n+i}\}^{-1}$ be the hazard sequence with respect to λ. Let t and n be related by $t \cdot \lambda_n = 1$. The next theorem specifies the rates of divergence of $E[V(t,\lambda)]$ in terms of n or h_n^{-1}.

Theorem 2. *For each t define $n(t)$ by $\lambda_{n(t)}^{-1} \leq t < \lambda_{n(t)+1}^{-1}$. Assume $\lambda_n \downarrow 0$.*

(i) *If $nh_n \to 0$, then*

$$(eh_{n(t)})^{-1} < E[V(t,\lambda)] \leq (h_{n(t)})^{-1}(1+o(1)).$$

(ii) *If nh_n is bounded away from 0, then there exists a finite constant C such that*

$$n(t)(1-e^{-1}) < E[V(t,\lambda)] \leq Cn(t).$$

Proof. For t of the form $t = 1/\lambda_n$ write

$$EV(t,\lambda) = \sum_{i=1}^{n} \left(1 - e^{-\lambda_i/\lambda_n}\right) + \sum_{i=1}^{\infty} \left(1 - e^{-\lambda_{n+i}/\lambda_n}\right). \tag{3.1}$$

The first term is at most n whereas by using the inequality $1-e^{-x} \leq x$, the second term is bounded by $(1/\lambda_n)\sum_{j=n+1}^{\infty} \lambda_j = h_n^{-1}$. The desired inequalities then follow easily. For general t the theorem also holds since $EV(t,\lambda)$ is continuous and monotone in t.

We shall show that the actual order is $O(n)$ for many interesting convergent series; but the order h_n^{-1} is also attainable.

Example 3.1. If $\lambda_n = n^{-p}$, $p > 1$, then for $i \leq n$,

$$\frac{\lambda_i}{\lambda_n} = \left(\frac{i}{n}\right)^{-p}$$

so the first term of (3.1) is $\sum_{i=1}^{n} f_p(i/n)$ where $f_p(x) = 1 - \exp\{-x^{-p}\}$. For $i \geq 1$,

$$\frac{\lambda_{n+i}}{\lambda_n} = \left(1 + \frac{i}{n}\right)^{-p}$$

and the second term of (3.1) is $\sum_{i=1}^{\infty} f_p(1+i/n)$. From (3.1), by using a Riemann sum approximation, we obtain for the sequence $\lambda_n = n^{-p}$,

$$\begin{aligned} E[V(t,\boldsymbol{\lambda})] &= n \cdot \int_0^\infty f_p(x)dx\{1 + o(1)\} \\ &\equiv t^{1/p} \cdot C_1(p)\{1 + o(1)\}, \end{aligned} \tag{3.2}$$

say. This is the case where the order $O(n)$ is attained. We remark that this relation is the classical species-area model of Arrhenius (1921). Herdan (1966, p.75) also indicated that this relation has been observed for many linguistic cases. Later, we shall show that $C_1(p) = \Gamma((p-1)/p)$.

Since $p > 1$ is required to ensure the convergence of $\sum n^{-p}$, we see that if there are infinitely many word types, the type-token relation cannot be of the order $O(t)$ or greater.

Example 3.2. Let $\lambda_n = n^{-1}(\log n)^{-2}$. Then $\sum \lambda_n$ converges, but barely. For this $\boldsymbol{\lambda}$, $h(n) \sim (n \log n)^{-1}$. By part (i) of Theorem 4 below, $E[V(t,\boldsymbol{\lambda})] = O(n \log n)$ $= O(h_n^{-1})$.

The two examples above exhibit the case where $\sum \lambda_i$ converges "slowly". Together they show that the type-token curve cannot be linear although it can be approached arbitrarily close from below. The next example is for the case that $\lambda_n \downarrow 0$ geometrically.

Example 3.3. If $\lambda_n = \alpha^n$, $0 < \alpha < 1$, we can derive the type-token curve somewhat differently. The basic idea is still (3.1), but for the geometric case $\lambda_i/\lambda_n = \alpha^{i-n}$, $\lambda_{n+i}/\lambda_n = \alpha^i$. Hence (3.1) reduces to

$$\begin{aligned} E[V(t,\boldsymbol{\lambda})] &= \sum_{i=0}^{n-1} (1 - \exp\{-\alpha^{-i}\}) + \sum_{i=1}^{\infty} (1 - \exp\{-\alpha^i\}) \\ &\equiv n - A_n(\alpha) + B(\alpha), \end{aligned}$$

where

$$A_n(\alpha) = \sum_{i=0}^{n-1} \exp\{-\alpha^{-i}\}, \quad B(\alpha) = \sum_{i=1}^{\infty}[1 - \exp\{-\alpha^{i}\}].$$

Note that since $0 < \alpha < 1$, both $A_n(\alpha)$ and $B(\alpha)$ are finite. Hence if $\lambda_n = \alpha^n$,

$$E[V(t,\boldsymbol{\lambda})] = -\log t/\log\alpha - A(\alpha) + B(\alpha) + o(1) \tag{3.3}$$

where

$$A(\alpha) = \lim_{n\to\infty} A_n(\alpha).$$

Note that (3.3) represents another well-known classical case: The Gleason (1922) species-area model.

In the following we shall give a simple method from which the type-token curve for various $\boldsymbol{\lambda}$ can be found. The idea is based on Theorem 3 below and the concept of regular variation (see Feller (1968, p.269) for the definition). It is possible to find $C_1(p)$ explicitly in Example 3.1. First consider the function

$$t\frac{d}{dt}E[V(t,\boldsymbol{\lambda})] = \sum_{i=1}^{\infty} \lambda_i t e^{-\lambda_i t},$$

which, on using the same technique as in Example 3.1, can be shown to be

$$n\int_0^\infty x^{-p}\exp\{-x^{-p}\}dx = n\Gamma((p-1)/p)/p.$$

Hence

$$\frac{t\frac{d}{dt}E[V(t,\boldsymbol{\lambda})]}{E[V(t,\boldsymbol{\lambda})]} \to \frac{\Gamma((p-1)/p)}{pC_1(p)} \equiv \rho,$$

say. But since $E[V(t,\boldsymbol{\lambda})] \sim C_1(p)t^{1/p}$, it is of regular variation with exponent $1/p$. By a result of von Mises (see Resnick (1987, p.21)), $\rho = 1/p$ so $C_1(p) = \Gamma((p-1)/p)$.

Theorem 3. *Assume for certain* $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots)$,

$$E[(V(t,\boldsymbol{\lambda})] = g(t)(1 + o(1))$$

for some regularly varying function g *with exponent* ρ. *If* $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots)$ *is another sequence of constants such that* $\mu_n/\lambda_n \to r > 0$, *then*

$$E[V(t,\boldsymbol{\mu})] = g(rt)(1 + o(1)).$$

Proof. For $\epsilon > 0$, let n_0 be chosen so that

$$(r-\epsilon)\lambda_n \le \mu_n \le (r+\epsilon)\lambda_n$$

if $n > n_0$. Then

$$\begin{aligned}
EV(t,\mu) &= \sum_{i=1}^{n_0}(1-\exp\{-\mu_i t\}) + \sum_{i=n_0+1}^{\infty}(1-\exp\{-\mu_i t\}) \\
&\le \sum_{i=1}^{n_0}(1-\exp\{-\mu_i t\}) + \sum_{i=n_0+1}^{\infty}(1-\exp\{-(r+\epsilon)\lambda_i\}) \\
&= \sum_{i=1}^{n_0}[\exp\{-(r+\epsilon)\lambda_i\} - \exp\{-\mu_i t\}] + \sum_{i=1}^{\infty}(1-\exp\{-(r+\epsilon)\lambda_i t\}).
\end{aligned}$$

On letting $t \to \infty$, the first summation tends to 0 and the second summation is of the form $g((r+\epsilon)t)(1+o(1))$. Since g varies regularly with exponent $\rho > 0$,

$$\begin{aligned}
g((r+\epsilon)t) &= g(rt)\cdot\left(1+\frac{\epsilon}{r}\right)^{\rho}\cdot\frac{L((r+\epsilon)t)}{L(rt)} \\
&= g(rt)\cdot\left(1+\frac{\epsilon}{r}\right)^{\rho}\cdot(1+o(1)).
\end{aligned}$$

A similar argument establishes the lower bound as $g(rt)(1-(\epsilon/r))^{\rho}(1+o(1))$. Since ϵ is arbitrary, this completes the proof.

4. The Variances

Theorem 5 below not only provides an easy vehicle to calculate the expected $V(t,\lambda)$, but also enables us to compute other moments. In this section we shall provide a general formula together with a few examples to illustrate the point. Write

$$\begin{aligned}
\mathrm{Var}[V(t,\lambda)] &= \sum_{i=1}^{\infty}\exp\{-\lambda_i t\}(1-\exp\{-\lambda_i t\}) \\
&= \sum_{i=1}^{\infty}[\exp\{-\lambda_i t\} - \exp\{-2\lambda_i t\}] \\
&= \sum_{i=1}^{\infty}[(1-\exp\{-2\lambda_i t\}) - (1-\exp\{-\lambda_i t\})] \\
&= E[V(2t,\lambda)] - E[V(t,\lambda)]. \qquad (4.1)
\end{aligned}$$

We see that for the case $\lambda_n = n^{-p}$, $p > 1$, (4.1) reduces to $(2^{1/p}-1)t^{1/p}C_1(p)(1+o(1))$ and C_2 of (1.1) is $(2^{1/p}-1)C_1(p)$. Whereas if $\lambda_n = \alpha^n$, (4.1) becomes

$-\log 2/\log\alpha + o(1)$. Hence C_3 of (1.2) is $-(\log\alpha)^{-1}$. This implies that for the geometric case

$$V(t,\lambda) = -\log t/\log\alpha + O_p(1). \tag{4.2}$$

The basic idea exhibited in (4.1) can be exploited further to find a general formula for all cumulants of $V(t,\lambda)$. Let I be a Bernoulli random variable $b(1;p)$, and let τ_k be its kth cumulant. Then $\tau_1 = p$ and $\tau_2 = p(1-p)$. In general, τ_k is a degree k polynomial in p, and also a degree k polynomial in q $(= 1-p)$. Write

$$\tau_k = a_{k0} + a_{k1}q + \cdots + a_{kk}q^k. \tag{4.3}$$

The values of these coefficients can be explicitly found. A rather tedious, but nevertheless straightforward, calculation shows that

$$a_{ki} = \sum_{r=i}^{k}\sum_{j=0}^{r}(-1)^{i+j-1}j^{k-1}r(r+1)\binom{r}{i}\binom{r-1}{j-1}. \tag{4.4}$$

It can be shown that for $k > 1$, $a_{k0} = 0$ and

$$a_{k1} + a_{k2} + \cdots + a_{kk} = 0. \tag{4.5}$$

Replacing q by $\exp\{-\lambda_i t\}$ in (4.3) and summing over i, we obtain the kth cumulant, τ_k, of $V(t,\lambda)$. Hence for $k > 1$,

$$\begin{aligned}\tau_k &= \sum_{i=1}^{\infty}\sum_{j=1}^{k} a_{kj}\exp\{-\lambda_i t\}\\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{k}[-a_{kj}(1-\exp\{-\lambda_i tj\}) + a_{kj}]\\ &= -\sum_{j=1}^{k} a_{kj}EV(jt,\lambda).\end{aligned} \tag{4.6}$$

For the special case that $\lambda_n = \alpha^n$, we may use (3.3) to find all cumulants of $V(t,\lambda)$ as $t \to \infty$. The final result is, for $k > 1$

$$\tau_k = (a_{k2}\log 2 + a_{k3}\log 3 + \cdots + a_{kk}\log k)/\log\alpha + o(1). \tag{4.7}$$

Finally, we remark here that it is easier to find the factorial cumulants of V (Kendall and Stuart (1977, p.77)).

5. Asymptotic Distribution of $V(t)$

We have seen in Section 4 that if $\lambda_n \downarrow 0$ geometrically, then the limit of $\mathrm{Var}[V(t,\lambda)]$ is finite. In this case, we do not need a normalizing rate for $V(t,\lambda)$ in order to obtain an asymptotic distribution. For other cases, e.g., $\lambda_n = n^{-p}$, the variance of $V(t,\lambda)$ tends to ∞. In this section, we show that if $\mathrm{Var}[V(t,\lambda)] \to \infty$, then it is possible to properly normalize $V(t,\lambda)$ to obtain a weak limit.

Write, as in (2.1), $t = \lambda_n^{-1}$ and

$$V(t,\lambda) = \sum_{j=1}^{n} I_{[X_j(t)\geq 1]} + \sum_{i=1}^{\infty} I_{[X_{n+i}(t)\geq 1]} \equiv V_1 + V_2,$$

say. Let $S_n' = \mathrm{Var}(V_1)$, $S_n'' = \mathrm{Var}(V_2)$, and let $S_n = S_n' + S_n''$.

Theorem 4. *If $S_n \to \infty$ then*

$$S_n^{-\frac{1}{2}}(V(t,\lambda) - EV(t,\lambda)) \xrightarrow{\mathcal{L}} N(0,1).$$

Proof. If $S_n \to \infty$, either $S_n' \to \infty$ or $S_n'' \to \infty$ or both. If $S_n' \to \infty$ then V_1 is a row-independent sum of indicator random variables and the Lindeberg condition (e.g. Chow and Teicher (1978, p.290)) holds. If $S_n'' \to \infty$, V_2 is also asymptotically normal by the same reason. Note that V_1 and V_2 are independent, and if $S_n' \to \infty$ faster than S_n'', then

$$S_n^{-\frac{1}{2}}(V(t,\lambda) - E[V(t,\lambda)]) \sim S_n^{-\frac{1}{2}}(V_1 - E[V_1]).$$

The other two cases are similar and the theorem is proved.

McNeil (1973) establishes a similar result, using several assumptions. Theorem 4 is more transparent and, in view of the Feller-Lindeberg condition, requires only the minimum assumption that $S_n \to \infty$.

To close this section, we provide a condition on λ, under which $S_n \to \infty$. The idea comes from the case where $\lambda_n = \alpha^n$ for some $\alpha \in (0,1)$. It turns out that the geometrical rate plays an important role.

Theorem 5. *If $\mathrm{Var}[V(t)] \leq C < \infty$ for all t, then there exists $\alpha \in (0,1)$ such that $\lambda_n \leq \alpha^n$ for all n sufficiently large.*

Proof. Let $f(t) = E[V(t)]$. Then, by (4.1),

$$\begin{aligned}\mathrm{Var}[V(t)] &= E[V(2t)] - E[V(t)]\\ &= f(2t) - f(t) \leq C\end{aligned}$$

for all t. By induction,

$$f(2^k t) - f(t) \leq kC.$$

Letting $2^k = x$ and $t = 1$, we have, for all x,

$$f(x) \leq f(1) + C_1 \log x \leq C_2 \log x \tag{5.1}$$

where C_1, C_2 are constants. Now we have

$$f(t) \geq \sum_{i=1}^{n} \left[1 - e^{-\lambda_i/\lambda_n}\right] \geq (1 - e^{-1})n. \tag{5.2}$$

By (5.1) and (5.2), there exists a constant C_3 such that

$$n \leq C_3 \log t = -C_3 \log \lambda_n.$$

Hence $\lambda_n \leq \alpha^n$ where $\alpha = \exp\{-1/C_3\} < 1$. This completes the proof.

Theorem 5 implies that if λ_n is not bounded above geometrically, then $S_n \to \infty$ and by Theorem 4, the central limit theorem holds for $V(t)$. It would be nice to show that if $\lambda_n \leq \alpha^n$ then $S_n < \infty$. However, despite its simplicity, we are unable to establish this result.

6. The Finite Vocabulary Case

We have so far considered only the case where $\lambda_n > 0$ for all n; i.e., the infinite vocabulary case. If $\lambda_n = 0$ for all $n > N$, then the limit of $V(t)$ is trivial (see Brainerd (1982, p.788)). Hence it is the intermediate case that we shall be interested in: that where both t and N are large but neither is infinite. Therefore, we shall be more concerned with the order of magnitude of $E[V(t)]$.

To fix ideas we shall only consider the case that $\lambda_n = n^{-p}$, $n = 1, 2, \ldots, N$; $\lambda_n = 0$, $n > N$. Admittedly, this is one of the cases that enables us to compute some of the relevant quantities. But we are hoping that this will provide a basis for comparison between the finite and infinite vocabulary cases.

Following Example 3.1, we first note that the $o(1)$ term of (3.2) is $O(1/n)$, where n and t are related by $t\lambda_n = 1$. We then follow the proof of Theorem 1, where the only change we need to make is to replace ∞ in (3.1) by $N - n$ to reflect the restriction that $\lambda_i = 0$ for $i > N$. To avoid notational confusion, λ_n in (3.1) is best replaced by the quantity n^{-p}, which is independent of the sequence $\boldsymbol{\lambda}$. Then (3.2) becomes

$$E[V(t, \boldsymbol{\lambda})] = n \cdot \int_0^{N/n} f_p(x)dx \left\{1 + O\left(\frac{1}{n}\right)\right\} \tag{6.1}$$

where $n = t^{1/p}$. On comparing (3.2) with (6.1), we see that the only difference is the finite-vocabulary correction N/n. As a simple check, we may let $t \to \infty$ (or

$n \to \infty$) in (6.1) and find its limit. By the L'Hôpital rule, it is easy to see that the limit of $E[V(t,\lambda)]$ is N, as it should. Let $g(t,N)$ denote the right-hand-side of (6.1). By (4.1), we have

$$\mathrm{Var}[V(t,\lambda)] = \{g(2t,N) - g(t,N)\} \cdot \{1 + O(t^{-1/p})\}. \tag{6.2}$$

We remark here that when $N < \infty$, there is no need to restrict $p > 1$. Hence (6.2) includes Zipf's case ($p = 1$).

The case where $\lambda_n = \alpha^n$ for $n < N$ only can be dealt with similarly, but more simply. Following the development of Example 3.3, it follows that, if $n \leq N$,

$$E[V(t,\lambda)] = n - A_n(\alpha) + B_{N-n}(\alpha) \tag{6.3}$$

where

$$A_n(\alpha) = \sum_{i=0}^{n-1} \exp\{-\alpha^{-i}\}$$

$$B_{N-n}(\alpha) = \sum_{i=1}^{N-n} [1 - \exp\{-\alpha^{i}\}],$$

and n and t are related by $t \cdot \alpha^n = 1$; whereas if $n > N$,

$$E[V(t,\lambda)] = N - \sum_{j=n-N}^{n-1} \exp\{-\alpha^{j}\}. \tag{6.4}$$

Combining (6.3) and (6.4) yields

$$E[V(t,\lambda)] = \min\{n, N\} + O(1).$$

Hence, by (4.1),

$$\mathrm{Var}[V(t,\lambda)] = O(1).$$

This holds true even if $N = \infty$. Note that for the case $N < \infty$ we do not need the restriction $\alpha < 1$.

If $\alpha = 1$, then both (6.3) and (6.4) reduce to

$$E[V(t,\lambda)] = N(1 - e^{-1})$$

but this derivation is wrong because for $\alpha = 1$, $t = 1/\alpha^n = 1$ cannot tend to ∞. We need to go back to (3.1) and find

$$E[V(t,\lambda)] = \sum_{t=1}^{N}(1 - e^{-t}) = N(1 - e^{-t}).$$

7. Conclusion

It has been shown that if we take Fisher's animal trapping setup the species area (or type token) curve can be "derived" from the structure of the arrival rates. In certain special cases specific functional forms can be found. The basic connection between λ_n and $EV(t)$ is Tauberian in nature, and the argument is elementary and no differential equation is needed.

Acknowledgement

The author thanks Sam Koltz and Joe Gani for valuable comments and he also thanks the referees. The research is partially supported by the National Science Council of the Republic of China.

References

Arrhenius, O. (1921). Species and area. *J. Ecology* **9**, 95–99.

Brainerd, B. (1982). On the relation between the type-token and species-area problems. *J. Appl. Probab.* **19**, 785–793.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory*. Springer-Verlag, New York.

Daley, D. J., Gani, J. M. and Ratkowsky, D. A. (1988). Markov chain models for type-token relationship. Technical Report 71, Program in Statistics and Applied Probability, University of California, Santa Barbara.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63**, 435–447.

Emigh, T. H. (1983). On the number of observed classes from a multinomial distribution. *Biometrics* **39**, 485–491.

Engen, S. (1978). *Stochastic Abundance Models*. Chapman and Hall, London.

Feller, W. (1968). *An Introduction of Probability Theory and Its Applications*, Vol. 2. John Wiley, New York.

Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecology* **12**, 42–58.

Gani, J. (1985). Literature and statistics. In *Encyclopedia of Statistical Sciences*, Vol. 5 (Edited by N. L. Johnson and S. Kotz), 90-95. John Wiley, New York.

Gleason, H. A., Sr. (1922). On the relation between species and area. *Ecology* **3**, 156–162.

Guiraud, P. (1959). *Problemes et Méthodes de la Statistique Linguistique*. D. Reidel, Dordrecht, Holland.

Herdan, G. (1960). *Type-token Mathematics: A Text Book of Mathematical Linguistics*. Mouton, The Hague.

Herdan, G. (1966). *The Advanced Theory of Language as Choice and Chance*. Springer-Verlag, New York.

Kendall, M. and Stuart A. (1977). *The Advanced Theory of Statistics*, Vol. 1. 4th edition. Griffin.

Lewin, R. (1983). Santa Rosalia was a goat. *Science* **221**, 636–639.

McNeil, D. R. (1973). Estimating an author's vocabulary. *J. Amer. Statist. Assoc.* **68**, 92–96.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes.* Springer-Verlag, New York.

Sichel, H. A. (1986). Word frenquency distributions and type-token characteristics. *Math. Sci.* **11**, 45–72.

Tipper, J. C. (1979). Rarefaction and rarefiction—the use and abuse of a method in paleoecology. *Paleobiology* **5**, 423–434.

Walton, G. S. (1986). The number of observed classes from a multiple hypergeometric distribution. *J. Amer. Statist. Assoc.* **81**, 169–171.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary.* Cambridge University Press.

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan

(Received April 1990; accepted June 1991)