*J. Appl. Prob.* **9**, 507–518 (1972)
*Printed in Israel*

# ON THE RELATION BETWEEN TYPES AND TOKENS IN LITERARY TEXT

BARRON BRAINERD, *University of Toronto*

**Abstract**

The ratio of the number $X_n$ of different words (types) in a text of length $n$ (token) words to $n$ has received considerable attention in the literature of statistical linguistics. The present note contains two stochastic models for $X_n$ based on an inhomogeneous discrete Markov process of the pure birth type where the transition probabilities take certain forms depending only upon $n$. These models are then tested against data obtained from the plays of William Shakespeare.

TYPE-COUNT; TOKEN-COUNT; SHAKESPEARE; INHOMOGENEOUS DISCRETE MARKOV PROCESS; VOCABULARY OF AN AUTHOR; LITERARY TEXT

## 1. Introduction

One of the first numerical studies of literary style was concerned with the relationship between the number of different words in a text (the *type-count*) and the total number of words (the *token-count*) in the text. Thomson and Thompson (1915) attempted to extrapolate, from a study of this relationship, the size of the lexicon from which a particular writer draws his vocabulary. A number of other authors have studied this relationship as an index of the style of particular writers: Muller (1965), Müller (1969), Herdan (1966), Yule (1944) to name but a few. Related studies with different theoretical points of departure and treating different, but related, problems are found in Simon ((1955), (1960)), Good ((1953), (1969)), J. B. Carroll's article in Kučera and Francis (1967) and Carroll (1968).

In the present note, we try to develop a stochastic model of this type–token relationship starting from the point of view that a literary text is a stochastic process. We will be concerned with the variate $X_n$, the number of different words in a text of length $n$ words. From a deterministic point of view, one might intuitively conceive of $X_n$ having the following properties:

(i) for very small $n$, $\Delta X_n = X_{n+1} - X_n$ is almost surely 1,

(ii) for all $n$, $\Delta X_n \geqq 0$,

(iii) for $n \to \infty$, $\lim X_n = M$ where $M$ might be conceived of as the maximal effective vocabulary of the writer in a single literary effort. (We will see that, if

Received in revised form 22 October 1971. This paper was prepared with the support of the National Research Council of Canada.

our models have any relevancy, this $M$ does not represent the writer's total effective vocabulary.)

In Section 2 we discuss the general stochastic process associated with $X_n$; in Section 3 we study some specific hypotheses concerning this stochastic process, and in Section 4 we consider some methods of estimating the parameters involved in the stochastic process for $X_n$. Finally, in Section 5 we consider some specific data and assess the reliability of the models discussed earlier.

## 2. The stochastic process

Consider a literary text $w_1 w_2 \cdots w_n$ which evolves in unit time jumps as the writer puts down words; we disregard punctuation, paragraphing, chapter headings, etc., in the case of a prose work, and punctuation, act headings, scene headings, stage directions, character headings, etc., in the case of drama.

For a given time $n$, after the author has written a total of $n$ word-tokens, let $X_n$ represent the number of different words in the text.

Let $P(X_n = i)$ be the probability that $i$ word-types are represented in $n$ word-tokens of the text and let

$$P(X_{n+1} = i + 1 \mid X_n = i) = f(n, i)$$

stand for the conditional probability that $w_{n+1}$ is a new word not already appearing among the words $w_1 w_2 \cdots w_n$ which contain $i$ word-types. If we assume that the probability that the $(n+1)$th word will be new depends only on the number $i$ of different words already present among the $n$ previous words, we effectively make the Markov assumption (Bailey (1964), pp. 38–39), and the random variables $X_1, X_2, \cdots$ constitute an inhomogeneous discrete Markov process of the pure birth type. From this assumption, we see that the Chapman-Kolmogorov equation

$$(1) \quad P(X_n = i) = P(X_{n-1} = i)(1 - f(n-1, i)) + P(X_{n-1} = i-1) f(n-1, i-1)$$

holds for $n > 1$ and $0 < i < n$.

Some boundary values need special mention:

$$(2) \qquad P(X_1 = k) = 1,$$

$$(3) \qquad P(X_n = 1) = P(X_{n-1} = 1)(1 - f(n-1, 1)), \qquad n > 1,$$

$$(4) \qquad P(X_n = n) = P(X_{n-1} = n-1) f(n-1, n-1), \qquad n > 1.$$

To obtain some knowledge of the nature of $P(X_n = i)$, let us construct its generating function:

$$G(n, x) = \sum_{i=0}^{n} P(X_n = i) x^i.$$

Equations (1)–(4) yield

$$\text{(5)}\qquad \begin{aligned} G(1,x) &= x, \\ G(n,x) &= \sum_{i=1}^{n-1} P(X_{n-1}=i)\,(1+(x-1)f(n-1,i))x^i \qquad (n>1). \end{aligned}$$

At this point, some hypothesis about the form of $f(n,i)$ must be made; it seems reasonable to assume that the dependence of $f(n,i)$ upon $i$ is minimal. Indeed, a writer does not keep count of the word-types he has already used in his text: in the long run at least, he is oblivious of whether the word he adds has already appeared or not. Thus let us assume[1] that $f(n,i)=g(n)$, independently of $i$. Equation (5) becomes

$$G(n,x)=(1+(x-1)\,g(n-1))G(n-1,x) \qquad (n>1).$$

Therefore, if we take $g(0)=1$, so that $G(1,x)=x$,

$$\text{(6)}\qquad G(n,x)=\prod_{j=0}^{n-1}(1+(x-1)g(j)).$$

From (6), noting that

$$\ln G(n,x)=\sum_{j=0}^{n-1}\ln(1+(x-1)g(j)),$$

we can obtain the first two moments of $X_n$ as

$$\text{(7)}\qquad E(X_n)=G'(n,1)=\sum_{j=0}^{n-1}g(j)$$

and

$$\text{(8)}\qquad \sigma^2_{X_n}=E(X_n)-\sum_{j=0}^{n-1}[g(j)]^2.$$

## 3. Some special hypotheses about $g(n)$

The simplest hypothesis about $g(n)$ consonant with conditions (i), (ii), and (iii) of Section 1 is that

$$\text{(9)}\qquad g(n)=e^{-\alpha n},$$

in which case

$$\text{(10)}\qquad G(n,x)=\prod_{i=0}^{n-1}(1+(x-1)e^{-\alpha i}),$$

with

$$\text{(11)}\qquad E(X_n)=\frac{1-e^{-\alpha n}}{1-e^{-\alpha}},$$

[1] Another possibility suggested by Donald McNeil is that $f(n,i)=a(N-i)$, where $N$ is the author's limiting vocabulary. This model would give rise to very different results and remains to be considered.

$$(12)\qquad \sigma^2_{X_n} = \frac{1-e^{-\alpha n}}{1-e^{-\alpha}} - \frac{1-e^{-2\alpha n}}{1-e^{-2\alpha}} = E(X_n)\ \frac{e^{-\alpha}(1-e^{-\alpha(n-1)})}{1+e^{-\alpha}}\,.$$

From (11) we see that the expected number of different words $E(X_n)$ in an $n$-word work tends to $1/(1-e^{-\alpha})$ as $n$ tends to infinity. If the parameter $\alpha$ is specific to a particular work, this model suggests that an author's vocabulary approaches a fixed maximum *for that work*, namely $1/(1-e^{-\alpha})$. An author's maximum potential vocabulary might vary with varying $\alpha$ at least from genre to genre, if not from work to work.

It is possible to carry the model at least one step farther in complexity. We might take into consideration the presence of grammatical or "empty" words such as "the", "a", "of" and "is", which tend to be included in every text of the given type with a nearly constant relative frequency. Suppose these word-types are $\omega_1, \omega_2, \cdots, \omega_m$ with relative frequencies $\alpha_1, \alpha_2, \cdots, \alpha_m$ respectively.

As an initial model, assume that the writer, when he is composing a text, selects, his tokens in turn from an urn containing a proportion $\alpha_i$ of the word-type $\omega_i$, and a proportion $\beta = 1 - \Sigma_{i=1}^m \alpha_i$ of blank tokens. If a blank is drawn assume that the probability that it represents a new word that is not in $\{\omega_1, \cdots, \omega_m\}$ and has not appeared already in the text is $e^{-\lambda n^*}$ where $n^*$ is the number of tokens (other than instances of $\omega_1, \omega_2, \cdots, \omega_m$) already in the text.

If $K$ is the class of types $\omega_1, \omega_2, \cdots, \omega_m$, then after the $n$th word-token has been added to the text,

$$\begin{aligned}
&P\,(\text{text contains } n_i \text{ tokens of } \omega_i \text{ and } n - \Sigma_{i=1}^m n_i \text{ tokens of types not in } K)\\
&\qquad = P(n_1, n_2, \cdots, n_m)\\
&\qquad = \binom{n}{n_1, n_2, \cdots, n_m} \alpha_1^{n_1} \cdots \alpha_m^{n_m} \left(1 - \sum_{i=1}^m \alpha_i\right)^{n-\Sigma^m{}_{=1} n_i}.
\end{aligned}$$

In addition,

$$\begin{aligned}
P(X_{n+1} = i+1 \mid X_n = i) &= \sum_{n_1,\cdots,n_m} P(1 \text{ new word is added} \mid n_1, \cdots, n_m) P(n_1, \cdots, n_m)\\
&= \sum_{n_1,\cdots,n_m} \left\{ \left[ \sum_{i=1}^m P(w_{n+1} \text{ is an instance of } \omega_i, \omega_i \text{ is not in the text} \mid n_1, \cdots, n_m) \right] \right.\\
&\qquad \left. + P(w_{n+1} \notin K \text{ and is new} \mid n_1, \cdots, n_m) \right\} P(n_1, \cdots, n_m)\\
&= \left( \sum_{n_1,\cdots,n_m} \sum_{i \text{ s.t.} n_i = 0} \alpha_i P(n_1, \cdots, n_m) \right)\\
&\qquad + \left(1 - \sum_{i=1}^m \alpha_i\right) \left(\alpha_1 + \cdots + \alpha_m + \left(1 - \sum_{i=1}^m \alpha_i\right) e^{-\lambda}\right)^n\\
&= \left( \sum_{i=1}^m (1-\alpha_i)^n \alpha_i \right) + \beta(1 - \beta + \beta e^{-\lambda})^n,
\end{aligned}$$

where $\beta = 1 - \Sigma_{i=1}^{m} \alpha_i$. Thus, for this model, we can write[2]

$$g(k) = \left( \sum_{i=1}^{m} (1-\alpha_i)^k \alpha_i \right) + \beta(1-\beta+\beta e^{-\lambda})^k, \tag{13}$$

so that

$$E(X_n) = m - \left( \sum_{i=1}^{m} (1-\alpha_i)^n \right) + \frac{1-(1-\beta+\beta e^{-\lambda})^n}{1-e^{-\lambda}}, \tag{14}$$

$$\begin{aligned} \sigma^2_{X_n} = {} & \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \frac{1-[(1-\alpha_i)(1-\alpha_j)]^n}{\alpha_i+\alpha_j-\alpha_i\alpha_j} \\ & + \beta \frac{1-(1-\beta+\beta e^{-\lambda})^{2n}}{2(1-e^{-\lambda})-\beta(1-e^{-\lambda})^2} \\ & + 2\beta \sum_{i=1}^{m} \alpha_i \frac{1-(1-\alpha_i)^n(1-\beta+\beta e^{-\lambda})^n}{1-(1-\alpha_i)(1-\beta+\beta e^{-\lambda})}. \end{aligned} \tag{15}$$

For moderately large $n$, say $n \geqq 2{,}000$, these simplify to

$$E(X_n) \sim m + \frac{1-(1-\beta+\beta e^{-\lambda})^n}{(1-e^{-\lambda})}, \tag{16}$$

$$\begin{aligned} \sigma^2_{X_n} \sim {} & \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\alpha_i\alpha_j}{\alpha_i+\alpha_j-\alpha_i\alpha_j} + \frac{\beta}{(1-e^{-\lambda})(2-\beta(1-e^{-\lambda}))} \\ & + 2\beta \sum_{i=1}^{m} \frac{\alpha_i}{1-(1-\alpha_i)(1-\beta+\beta e^{-\lambda})}. \end{aligned} \tag{17}$$

## 4. Estimation of parameters

Let us return to the general case represented by Equations (6), (7), and (8). In its full generality, the situation involves making certain hypotheses about the form of $g(n)$, say

$$g(n) = h(n; \gamma_1, \cdots, \gamma_k),$$

where $h$ is a function of $n$ and, say $k$ parameters $\gamma_i$; we should then try to estimate the $\gamma_i$ using data from a given writer and genre. The accumulation of data for large values of $n$ poses some problems. Ideally, we would like to sample $X_n$ from a number of independent works $W_1, W_2, \cdots, W_p$ in a particular genre by a particular author at a number of places $n_1, n_2, \cdots, n_q$ in each work, so that we could obtain the number of types $x_{n_{i.j}}$ among the first $n_i$ tokens of $W_j$, and with these sample values estimate the various parameters. However, considering the magnitude of

[2] Annette Dobson has suggested that the $\alpha_i$'s might follow some distribution such as Zipf's law (Herdan (1966)). This would cut down the parameters to be estimated from $m + 3$ to a more manageable number.

the values of the $n_i$ necessary to use Equation (16) for example, such a study could be too costly to be made for this purpose alone.

Nevertheless, concordances often contain sufficient information to yield a large number of pairs $(n, X_n)$ for certain more or less random values of $n$, so that with the aid of certain regression models, we might attempt to estimate the parameters.

Assuming that the variation in the values $g(k)$ with $k$ is small over short intervals,[3] it seems reasonable to hypothesize that locally $X_n$ can be written

$$X_n = an + b + E_n, \tag{18}$$

where $E_n$ is an error random variable with zero mean and $a$ and $b$ are numerical constants. Then from Equation (7), we obtain

$$E(X_{n+1}) - E(X_n) = g(n) = a. \tag{19}$$

Given a sample of token-type pairs, the constants $a$ and $b$ can be estimated using regression theory so that near the mean point $(\bar{n}, \bar{X}_n)$ of the sample, we might expect to have $g(n) = a$.

Another m϶thod of estimating the parameters involved in $E(X_n)$ might be to divide the sample points into clusters $C_j$ and use their centres of gravity $(\bar{n}_j, \bar{X}_{nj})$ to estimate these parameters; or perhaps we might use the results of adjacent clusters $C_j$ and $C_{j+1}$ to estimate $g(\frac{1}{2}(\bar{n}_j + \bar{n}_{j+1}))$.

None of these methods are anything more than very *ad hoc* attempts to estimate the parameters. The problem of obtaining statistically adequate estimates is yet to be solved. However, since the problem of obtaining these statistically adequate estimates depends heavily on the stochastic model chosen, we have proceeded with the methods outlined above in order to obtain a rough test of the adequacy of the model under discussion.

In the following section each of the methods is used upon data taken from Shakespeare's comedies.

## 5. An illustrative example

From a concordance of the works of William Shakespeare, we obtain the information, Spevack (1968), given in Table 1 concerning the type- and token-counts for his comedies. A regression analysis of this data yields

$$X_n = (0.1060)n + 1116.59$$

as the best least squares fit with a correlation coefficient of $\rho = 0.82$ and with $\bar{X}_n = 3211.86$ and $\bar{n} = 19763.93$. We can use these values to obtain an estimate of $e^{-\alpha}$ from (9) and (19):

[3] A plot of the data given in the tables shows that this working hypothesis is not entirely unreasonable.

TABLE 1

| Play | $n$<br>No. of Tokens | $X_n$<br>No. of types | I | II | III | IV | V |
|---|---|---|---|---|---|---|---|
| Comedy of Errors | 14369 | 2522 | 7084.13 | 3257 | 1924.95 | 2872.13 | 2764.54 |
| The Tempest | 16036 | 3149 |  | 3274 |  | 3035.49 | 2924.63 |
| A Midsummer Night's Dream | 16087 | 2984 |  |  |  | 3040.16 | 2929.21 |
| Two Gentlemen of Verona | 16883 | 2718 |  |  | 1925.75 | 3110.59 | 2998.62 |
| Twelfth Night | 19401 | 3096 |  | 3290 |  | 3306.29 | 3192.80 |
| The Taming of the Shrew | 20411 | 3240 | 7833.92 |  |  | 3374.46 | 3260.97 |
| Much Ado about Nothing | 20768 | 2954 |  |  | 1926.01 | 3397.28 | 3283.85 |
| Merchant of Venice | 20921 | 3265 |  |  |  |  | 3293.48 |
| Love's Labour's Lost | 21033 | 3772 |  | 3294 |  | 3413.81 | 3300.45 |
| Merry Wives of Windsor | 21119 | 3267 |  |  |  |  | 3305.77 |
| Measure for Measure | 21269 | 3325 |  | 3294 |  |  | 3314.96 |
| As You Like It | 21305 | 3248 |  |  |  |  | 3317.15 |
| All's Well That Ends Well | 22550 | 3513 |  | 3296 |  | 3502.03 | 3389.38 |
| A Winter's Tale | 24543 | 3913 | 8264.17 | 3298 | 1926.05 | 3602.98 | 3491.86 |

$$(e^{-\alpha})^{\wedge} = 0.99988645.$$

The estimates of $E(X_n)$ for the various comedies using (11) are given in column I of Table 1. They are far too large, so we look at another method of estimation for the same model.[4]

Using the cluster technique, we consider those comedies with $20411 \leqq n \leqq 21305$ and obtain their centre of gravity

$$(\bar{n}, \bar{X}_n) = (20975.14, 3295.86).$$

Newton's method and Equation (11) yield the estimate

$$(e^{-\alpha})^{\wedge} = 0.99969693.$$

Some of the values of $E(X_n)$ using this estimate are given in column II of Table 1. Again they are too large in the lower part of the scale, while in the upper part they tend to be too small.

Assume for the present that the value of $X_n$ in a given sample is not a matter of conscious choice by the writer, so that it is independent of whether the text of length $n$ used is continuous or not. Then we are at liberty to use the speech of individual characters within a play to obtain our sample pairs $(n, X_n)$. If we look at the 45 characters (in all the comedies) with more than 2000 words of dialogue, and obtain the corresponding regression line, we find

$$X_n = (0.2277)n + 239.80$$

with $\rho = 0.97$ and $\bar{n} = 2849.29$. The estimate of $e^{-\alpha}$, in this case, is

$$(e^{-\alpha})^{\wedge} = 0.99948080.$$

The values of $E(X_n)$ for this estimate, given in column III of Table 1, are much too small, and again the rate of growth of $E(X_n)$ is too small. Thus the simple exponential model seems to be unsuitable, and we turn our attention to the mixed model given by Expression (13).

Using (13) as our model, we can obtain estimates for $\beta$, $m$, and $e^{-\lambda}$ as follows: the comedies with $20{,}411 \leqq n \leqq 21{,}305$ and those with $n = 16036$, 16087 and 16883 form natural groupings with $(\bar{n}, \bar{X}_n)$ equalling respectively $P_1 = (20975.14, 3295.86)$ and $P_2 = (16335.33, 2950.33)$. In addition, under the assumption of textual homogeneity mentioned above, we obtain a third mean sample point $P_3 = (2378.24, 779.38)$ for the 30 characters in the comedies which have between 2000 and 3000 words of dialogue. If we connect these three points with line segments $P_3P_2$ and $P_2P_1$ and use their slopes as estimates of $g(n)$ for $n$ equal to the abscissa of their mid-point in each case, we obtain $g(9357) = 0.1555$ and

[4] Some non-linear regression models we tried resulted in estimates of $e^{-\alpha}$ similar to that obtained above. In particular, $X_n = an^b$ yielded $(e^{-\alpha})^{\wedge} = 0.99988432$, $X_n/n = ae^{bn}$ yielded $(e^{-\alpha})^{\wedge} = 0.99988362$, and $X_n = a + bn + cn^2$ yielded $(e^{-\alpha})^{\wedge} = 0.99989033$.

$g(18655) = 0.0745$, assuming that $n$ is large enough to render the first term of (13) negligible. In this way we obtain the estimates

$$\hat{q} = (1 - \beta + \beta e^{-\lambda}) = 0.99992083,$$

$\hat{\beta} = 0.33$, and $(e^{-\lambda})^\wedge = 0.99975736$. Using $\bar{n} = 16335.33$ in (17), we find that for $\bar{X}_n = (E(X_{\bar{n}}))^\wedge = 2950.33$,

$$\hat{m} = \bar{X}_n - \frac{1 - \hat{q}^{\bar{n}}}{1 - (e^{-\lambda})^\wedge} = -40.54,$$

which is clearly inadmissible in terms of the model. However if we use $P_3$ as an estimator, we obtain $\hat{m} = 72.07$ which is more acceptable. We return later to the question of whether it is reasonable that 72 types account for 67% of the tokens.

Let us now compute some of the expected values for various values of $n$ in Table 1. These are given in column IV, and appear to provide a superior fit to that obtained using the other model. However, if we refine our methods of estimation, we can obtain even better results.

If we use three clusters of data with mean token-counts $\bar{n}_i$ $(i = 1,2,3)$ and corresponding mean type-counts $\bar{X}_{n,i}$, then we can use the equations

$$E(\bar{X}_{n,i}) = m + \frac{1 - q^{\bar{n}_i}}{(1 - e^{-\lambda})} \qquad (i = 1,2,3) \tag{20}$$

to estimate the parameters involved. In this case, we can write

$$\frac{E(\bar{X}_{n,1}) - E(\bar{X}_{n,3})}{E(\bar{X}_{n,1}) - E(\bar{X}_{n,2})} = \frac{q^{\bar{n}_3} - q^{\bar{n}_1}}{q^{\bar{n}_2} - q^{\bar{n}_1}}, \tag{21}$$

and if we try our previously obtained value $\hat{q}_0 = 0.99992083$ as a first approximation, then for the data points $P_1, P_2, P_3$, Newton's method yields the estimate $\hat{q} = 0.99992511$. Since

$$1 - e^{-\lambda} = \frac{1 - q}{\beta}, \tag{22}$$

and

$$\hat{\beta} = \frac{(E(\bar{X}_{n,1}) - E(\bar{X}_{n,i}))(1 - \hat{q})}{\hat{q}^{\bar{n}_i} - \hat{q}^{\,\bar{n}_1}}, \tag{23}$$

we find that $\hat{\beta} = 0.30$, $(e^{-\lambda})^\wedge = 0.99975005$, and $\hat{m} = 127.63$.

The expected token-counts relative to the estimates just obtained appear in column V of Table 1. The agreement is clearly improved over column IV.

Insofar as the proposed model and these results are to be taken seriously, certain conclusions can be drawn from them:

(i) $\hat{m} = 127$ types account for approximately 70% of the text,

(ii) the upper limit on Shakespeare's possible vocabulary for a single comedy is

$$\lim_{n\to\infty} E(X_n) = m + \frac{1}{1-e^{-\lambda}} \sim 4127 \text{ words.}$$

Conclusion (i) does not seem entirely extraordinary: for in Kučera and Francis (1967), we find a much more diffuse type-token sample where the 127 most frequently used words account for about 50% of the total text obtained.

Conclusion (ii) must be considered a mark against the model if we think of lim $E(X_n)$ as an indication of Shakespeare's total vocabulary; for in the non-dramatic works there is a total of $n = 47824$ tokens and $X_n = 6797$ types. However, there may be reasons, not yet apparent, for the smaller limiting value of Shakespeare's vocabulary for a single comedy as compared with his total vocabulary.

Since we have no statistical control over our estimates, we will try this final method of estimation on other data: the 14 tragedies[5] and the 10 historical plays of Shakespeare, given in Spevack (1968). We list the former in Table 2 and the latter in Table 3.

TABLE 2

| Play | $n$ | $X_n$ | I |
|---|---|---|---|
| Macbeth | 16,436 | 3306 | 2939.82 |
| Pericles | 17,723 | 3270 | 3101.10 |
| Timon of Athens | 17,748 | 3269 | 3104.18 |
| Julius Caesar | 19,110 | 2867 | 3268.94 |
| Titus Andronicus | 19,790 | 3397 | 3349.03 |
| Two Noble Kinsmen | 23,403 | 3895 | 3751.51 |
| Anthony and Cleopatra | 23,742 | 3906 | 3787.36 |
| Romeo and Juliet | 23,913 | 3707 | 3805.33 |
| King Lear | 25,221 | 4166 | 3940.11 |
| Troilus and Cressida | 25,516 | 4251 | 3969.88 |
| Othello | 25,887 | 3783 | 4006.99 |
| Coriolanus | 26,579 | 4015 | 4075.26 |
| Cymbeline | 26,778 | 4260 | 4094.66 |
| Hamlet | 29,551 | 4700 | 4354.81 |

For the tragedies $n$ falls into two clusters, those plays with $23{,}403 \leqq n \leqq 26{,}778$ and those with $16{,}436 \leqq n \leqq 19{,}790$, with *Hamlet* a straggler at $n = 29{,}551$. To obtain a third cluster, we consider those tragedy characters with between 2000 and 3000 words of dialogue. From these three groups, we obtain the sample centres of gravity $(\bar{n}, \bar{X}_n)$:

[5] We have not included *Sir Thomas More* because it is too short for our approximations to take effect.

$$P_1 = (25129.88,\ 3997.88),$$
$$P_2 = (18161.40,\ 3221.80),$$
$$P_3 = (2443.08,\ 854.05).$$

Equations (20)–(23) yield the estimates

$$\hat{q} = 0.99997387,\ \hat{\beta} = 0.20,\ (e^{-\lambda})^{\wedge} = 0.99986870,\ \hat{m} = 323.65.$$

Column I of Table 2 shows the corresponding values of $E(X_n)$. Again note that the fit is reasonable but not as good as the comedy-fit.

Grouping the historical plays in a similar fashion, we obtain for the six plays with $23{,}295 \leqq n \leqq 25{,}706$, $P_1 = (24384.67, 3949.67)$, for the three plays with $20{,}386 \leqq n \leqq 21{,}809$, $P_2 = (20903.33, 3686.33)$, and for the history characters with between 2000 and 3000 words of speech, we obtain $P_3 = (2463.59, 867.65)$. These points yield the estimates

$$\hat{q} = 0.99994023,\quad \hat{\beta} = 0.29,\quad (e^{-\lambda})^{\wedge} = 0.99979550,\quad \hat{m} = 198.08.$$

These estimates yield the values of $E(X_n)$ given in column I of Table 3. They also constitute a reasonably good fit of the data.

Table 3

| Play | $n$ | $X_n$ | I |
|---|---|---|---|
| King John | 20,386 | 3576 | 3642.29 |
| Henry VI Part 1 | 20,515 | 3812 | 3653.40 |
| Richard II | 21,809 | 3671 | 3760.19 |
| Henry VI Part 3 | 23,295 | 3581 | 3873.05 |
| Henry VIII | 23,325 | 3558 | 3875.23 |
| Henry IV Part 1 | 23,955 | 3817 | 3920.05 |
| Henry VI Part 2 | 24,450 | 4058 | 3940.47 |
| Henry V | 25,577 | 4562 | 4027.98 |
| Henry IV Part 2 | 25,706 | 4122 | 4036.12 |
| Richard III | 28,309 | 4092 | 4187.71 |

The variation in $\hat{m}$ is not unexpected, for the first 324 of the ranked words in the Kučera and Francis (1967) list account for approximately 58% of the total text. Comparing this with earlier remark about $\hat{m} = 127$, we see that it is not unreasonable that the estimates of $m$ react strongly to small variations in $\hat{\beta}$.

In summary, it appears that the simple exponential model yields a poor fit, while the mixed model yields a fairly serviceable one. Perhaps the hypothesis of more than one body of types whose use decays exponentially with $n$ might provide for a yet more successful fit.

Among the admittedly *ad hoc* methods of estimation those involving the linear regression are least reliable, with polygonal estimates of the slope being somewhat better. The best method appears to be the conceptually simplest but computationally most arduous—that of using the centres of gravity of clusters of data points.

Of some interest are the values of $\lim E(X_n)$ obtained using the second model. For Comedies, Histories, and Tragedies these values are 4127, 5088, and 7940 respectively. They appear to increase with the seriousness of the genre. A search for an explanation of this phenomenon might prove rewarding.

## Acknowledgment

The author wishes to thank D. B. Delury, J. C. Ogilvie, G. S. Watson, Annette J. Dobson and Donald McNeil for their helpful suggestions.

## References

BAILEY, N. T. J. (1964) *The Elements of Stochastic Processes*. John Wiley and Sons, New York.

CARROLL, J. B. (1968) Word-frequency studies and the log-normal distribution. In *Proc. Conf. on Language and Language Behavior* (G. M. Zale, Ed.) Appleton-Century-Croft, New York.

GOOD, I. J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* **40,** 237–264.

GOOD, I. J. (1969) Statistics of Language. *Encyclopedia of Linguistics, Information, and Control.* (A. R. Meetham, Ed.) Pergamon Press, London, pp. 567–581.

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

KUČERA, H. AND FRANCIS, W. N. (1967) *Computational Analysis of Present-Day American English*. Providence, R. I.

MULLER, C. (1965) Lexical distribution reconsidered: the Waring-Herdan formula. First appeared in French in *Cahiers de Lexicologie*, Vol. 6 and reprinted in *Statistics and Style*, L. Doležel and R. W. Bailey, Editors. American Elsevier, New York 1969.

MÜLLER, W. (1969) Gedanken zur automatischen Analyse von Normen und Normabweichungen. *Muttersprache* **79,** 301–304.

SIMON, H. A. (1955) On a class of skew distribution functions. *Biometrika* **42,** 425–439.

SIMON, H. A. (1960) Some further notes on a class of skew distribution functions. *Information and Control* **3,** 80–88.

SPEVACK, M. (1968) *A Complete and Systematic Concordance to the Works of Shakespeare*, vols. I–VI. George Olms, Hildesheim.

THOMSON, G. H. AND THOMPSON, J. R. (1915) Outlines of a method for the quantitative analysis of writing vocabularies. *British J. Psychology* **8,** 52–69.

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