Word Length Frequency and Distribution in English: Part I. Prose

Observations, Theory, and Implications

Hideaki Aoyama and John Constable
Faculty of Integrated Human Studies, Kyoto University, Japan

Abstract

Recent observations in the theory of verse and empirical metrics have suggested that constructing a verse line involves a pattern-matching search through a source text, and that the number of found elements (complete words totalling a specified number of syllables) is given by dividing the total number of words by the mean number of syllables per word in the source text. This paper makes the latter point explicit mathematically, and in the course of this demonstration shows that the word length frequency totals in English prose output are distributed geometrically (previous researchers reported an adjusted Poisson distribution), and that the sequential distribution is random at the global level, with significant non-randomness in the fine structure. Data from a corpus of just under two million words and a syllable-count lexicon of 71,000 word forms is reported, together with some speculations concerning the relationship between the word length frequency distributions in output and in the lexicon. The pattern-matching theory is shown to be internally coherent, and it is observed that some of the analytical techniques described here form a satisfactory test for regular (isometric) lineation in a text.

1 Introduction

The making of isometric verse lines is a pattern-matching exercise, and in this respect contrasts sharply with the generation of ordinary language output. The composer of verse must first generate source language, and then search through it for elements which form, or may be accumulated to form, the desired pattern. This pattern is usually specified in relation to a selection from the available surface features of the language, those in English verse, for example, being based principally on the patterning of stressed and unstressed syllables to form arrangements of beats and offbeats (Attridge, 1982, 1995). In earlier work, one of us (Constable, 1997) has observed that these rules lead to the implication of a simpler and more readily studied rule, of which a composer is usually unaware, namely that a lineated text must be a sequence of lines consisting of

Correspondence:
Hideaki Aoyama,
Faculty of Integrated Human Studies, Kyoto University,
Kyoto, 606-8501, Japan
E-mail:
aoyama@phys.h.kyoto-u.ac.jp

complete words totalling *n* syllables, where *n* can be a range. This may be formulated as follows:

> 'In every consecutive section of *n* syllables there must only be complete words.'

The composition of word strings to fit the lines defined by a rule of this type involves finding sequences of complete words totalling *n* syllables, or constructing them from sequences of less than *n* syllables. In either case, the activity is a pattern-matching search for a target, and the frequency of the target in the source language matters to the composer, since a smaller number of found elements will restrict more severely the communicative options open. That is to say, the greater the number of targets in the source text, the more probable it is that the composer will find pieces that function adequately with regard to his or her purpose. Therefore, in order to facilitate line composition, authors are expected to take whatever action they can to increase the frequency of target elements. An effective way of doing this is to reduce the mean number of syllables per word in the source language being searched. Recent work in empirical metrics (Constable, 1997, p. 182) has demonstrated the relationship between mean word length and target frequency empirically from small samples of English prose, and has claimed that the average frequency of a target object (a sequence of words totalling *n* syllables) is given by the total number of words divided by the mean number of syllables per word. The mathematical reformulation required to substantiate and explain the factors underlying this effect involves detailed observations regarding the frequency and sequential word length distributions typical of English, and we will now turn to this task.

2 Methodology: Global Structures and Fine Structures

In approaching word length data, it is of crucial importance to form an analytical strategy suitable for the purpose in hand. Given a distribution, either of frequencies or probabilities, fitting it with a function with any degree of accuracy is a straightforward matter, provided that one has a large number of functions, each with a large enough number of parameters. Previous research on word length frequency undertaken by the Gottingen Word Length Project (see Best and Altmann, 1996, for an overview of the project and a bibliography; prominent papers in the project include: Wimmer and Altmann, 1994; Becker, 1996; Dittrich, 1996; Frischen, 1996; Riedemann, 1996; Röttget, 1996; Wimmer *et al.*, 1996; Ziegler, 1996; Zuse, 1996) has collected data relating to word length frequencies, usually from rather small samples, and then used software, the Altmann Fitter, to compute a best fit description, which in most cases proves to be an adjusted Poisson distribution (Best and Altmann, 1996). Such a fitting procedure is guaranteed to work: word forms of more than seven syllables are not numerous in the lexicon, and they are very

infrequent in output, so there are usually only seven data points. Therefore, an accurate fit may be found with little difficulty if we have several hundred functions, each with several parameters, from which to choose. Mathematically, if there are seven data items, a general function with seven parameters would suffice for a *perfect* fit.

However, a procedure of this kind is quite irrelevant to the analysis in hand and, although such a fit might serve some technical purposes, it brings no insight into or understanding of the nature of the language data under consideration. Abstraction and understanding will only result when we can find a fit with a much smaller number of parameters than that of the data. Any deviation from this simple ansatz should be regarded as a subtle variation from the basic finding.

Our research is directed with this general analytical principle in mind, and we have therefore chosen not to fit the distributions using a large set of functions and parameters. Rather, we will first extract a simple property that describes the overall, global structure of the data, from which we aim to obtain a deeper knowledge of the English language, if only in its syllabic organization. With this in hand, we can then proceed to the study of any deviations from this global structure which might be apparent in the fine structure of the data.

2.1 The corpus

We have analysed the word-length structure of almost two million words of prose by various authors as listed in Table 1. The texts were chosen with no other view than convenience, Constable having marked them in the course of other research. We acknowledge that a more principled choice would be desirable, and indeed hope to undertake such work ourselves, but we believe that this corpus is sufficient for present purposes. The syllabic data were obtained by using a simple marking program, constructed by Constable, which reads text and uses a custom-built lexicon to determine the syllabic count of each word form.

Table 1 Content of the corpus; authors, titles of the sources, and number of words from each author

Author	Title(s)	Words
Bunyan, John	*Pilgrim's Progress*	52,504
Eliot, George	*Middlemarch*	317,827
Frankau, Gilbert	*Woman of the Horizon* (section, first 67 pages 10 chapters)	24,597
Goldsmith, Oliver	*Vicar of Wakefield*	63,096
James, Henry	*The Altar of the Dead*; *The Ambassadors*; *The American*, *The Aspern Papers*, *Confidence*; *Daisy Miller*; *Death of the Lion*; *The Europeans*; *The Figure in the Carpet*; *The Golden Bowl*; *An International Episode*; *Portrait of a Lady*; *Roderick Hudson*; *Sacred Fount*; *Turn of the Screw*; *Watch and Ward*, *Washington Square*	1,285,041
Kipling, Rudyard	*Rewards and Fairies*; *The Jungle Book*	115,602
Milton, John	*History of Britain*; *Colasterion*; *Martin Bucer*	119,009
Total		1,977,676

When new word forms are encountered, the program requests human intervention, and the form is added to the lexicon. Consistency of syllable counting was ensured by the fact that only one user (Constable) has been responsible for building the lexicon, which at the time of writing contains 71,666 items. Abbreviations were expanded, and numbers were counted as if they were pronounced as hyphenated words (1,920 = one-thousand-nine-hundred-and-twenty), with the exception of years and dates which were counted in their normal pronounced form (1920 = nineteen twenty). None of these categories is, as it happens, frequent in our corpus.

Sections 2.2 to 2.4 below offer a technical and mathematical analysis of our results. Briefly, we show that (i) the word length frequency totals, i.e. the number of words of one syllable, two syllables, and so on, form a geometric series, and (ii) the word length sequence of English is random, in other words that the length of a word does not affect the length of preceding or succeeding words. Combined, these observations yield the proposition that the distribution of spaces in English text is random. Moreover, we show that the geometric distribution and the random sequencing of word length together account for the observation made by Constable (1997, p. 182) that the number of sequences of complete words totalling a specified number of syllables in a text could be calculated by dividing the total number of words by the mean number of syllables per word in that text.

2.2 Frequency and probabilities

In order to present the method of analysis in definite terms, we will first introduce our mathematical notations. We denote the syllabic data obtained by a series of integers N_i (i = 1, 2, 3, . . . I), where N_i is the number of syllables in the i-th word and I the total number of words in the data.

In the previous section we introduced the line definition rule, which specifies that for a lineated text 'every consecutive section of n syllables [. . .] must only be complete words'. The number of such sequences of words in a text for any given value of n can be computed easily. Say that we take a passage of prose and represent it as a series of numbers representing the number of syllables in each word:

Today	is	a	holiday	and	the	weather	is	bad
2	1	1	3	1	1	2	1	1

We could tally the frequency of word sequences of varying lengths as follows: 'Today' (two syllables), 'Today is' (three syllables), 'Today is a' (four syllables), 'Today is a holiday' (seven syllables), 'Today is a holiday and' (eight syllables), 'Today is a holiday and the' (nine syllables), 'Today is a holiday and the weather' (eleven syllables), 'Today is a holiday and the weather is' (twelve syllables), 'Today is a holiday and the weather is bad' (thirteen syllables). We would then move to the second word in the series and count forward from that position: 'is' (one syllable), 'is a' (two syllables), 'is a holiday' (five syllables), and so on. The process would

be repeated for all subsequent words. Such a method is somewhat misleading, since the text used is finite, and this causes an under-representation of longer elements in the final tally scores. However, we can solve this problem by treating the text as circular. With such a model, the last word of the original text is the last starting word for our counting, but we can count forward into the repeated words: 'bad' (one syllable), 'bad. Today' (three syllables), and so on.[1] A short text of this kind does not produce interesting results, but longer texts do. For example, if we take the first 1,007 words of Jane Austen's *Pride and Prejudice*, we find that it contains 756 sequences of complete words totalling one syllable, 752 sequences totalling two syllables, 757 totalling three syllables, and so on, all subsequent totals being similar. The mean of the values for lengths up to twenty syllables proves to be 757, which can be closely approximated to by dividing the total number of words (1007) by the mean number of syllables per word (1.33).

This procedure, though convenient as a introductory explanation, is very clumsy, and we can perform the operation rather more elegantly by observing that the number of sequences of words in a series N_i which would satisfy the line definition can be represented as follows:

$$n = \sum_{i=l}^{m \,(\mathrm{mod}\, l)} N_i. \tag{1}$$

In other words, if this equation is met, it means that the sequence of the words starting at the l-th word and ending at m-th words (or, $m - I$ word in case m exceeds I, $m - 2I$ if m exceeds $2I$, and so forth) satisfy the line definition rule for n-syllables.[1]

We define $L_{n,k}$ to be the number of occurrences that k sequential words satisfy the n-syllable line definition rule. It is straightforward to count this number $L_{n,k}$: the numbers $L_{n,1}$ are obtained simply by counting the numbers equal to n among the series $(N_1, N_2, \ldots, N_I)$. Next, the numbers $L_{n,2}$ are obtained by counting similarly for $(N_1 + N_2 + N_3, \ldots, N_I + N_1)$, $L_{n,3}$ from $(N_1 + N_2 + N_3, N_2 + N_3 + N_4, \ldots, N_I + N_1 + N_2)$, and so forth. By definition, the following identity is satisfied:

$$\sum_{n=1}^{\infty} L_{n,k} = I. \tag{2}$$

Since there are no zero-syllable words in English,

$$L_{n,k} = 0 \text{ if } n < k. \tag{3}$$

The quantity we are interested in is the number of sequences matching the line definition rule for *any* number of words, which is given by the following:

$$L_n = \sum_{k=1}^{n} L_{n,k}. \tag{4}$$

1 The upper limit of this sum implies that the line definition rule is applied with a 'periodic boundary condition', namely the data is treated as a circle by connecting the end of data with the beginning. As will become clear, this technical definition is justified by its utility in the following mathematical treatment. Alternatively, one could use a Dirichlet boundary condition, in which one simply terminates the data sequence at $i = I$. These boundary conditions, however, do not significantly affect the results as long as the data size is large, which is true for all the material we have analysed.

This counting algorithm has been coded in *Mathematica* by Aoyama, and is of course rather faster than the original algorithm described above and used by Constable (1997, p. 181). In Table 2, we give a partial list of $L_{n,k}$ and L_n obtained for all the data listed in Table 1.

For theoretical reasons it is best to deal with these quantities in a way which is independent of the data size. Therefore, we introduce the following normalized quantities:

$$P_{n,k} \equiv \frac{L_{n,k}}{I}, \quad Q_n \equiv \frac{L_n}{I}. \tag{5}$$

Due to the identity (Equation 2), the following is satisfied:

$$\sum_{n=1}^{\infty} P_{n,k} = 1. \tag{6}$$

Table 2 Number of strings $L_{n,k}$ and L_n that satisfy the n-syllable line definition rule for $n = 1$–30 and $k = 1$–5

n	$L_{n,1}$	$L_{n,2}$	$L_{n,3}$	$L_{n,4}$	$L_{n,5}$	L_n
1	1,433,426	0	0	0	0	1,433,426
2	371,500	1,025,719	0	0	0	1,397,219
3	122,179	558,679	733,202	0	0	1,414,060
4	40 314	246,132	611,737	531,686	0	1,429,869
5	9,048	99,647	348,154	583,554	387,684	1,428,087
6	1,082	33,891	169,801	411,842	524,656	1,425,780
7	119	9,790	73,374	238,242	439,613	1,426,115
8	6	2,983	27,502	121,820	293,091	1,426,874
9	2	660	9,832	54,843	171,197	1,426,456
10	0	146	2,902	22,952	88,979	1,426,660
11	0	26	878	8,329	42,419	1,426,218
12	0	3	219	3,012	18,275	1,426,323
13	0	0	60	983	7,464	1,426,066
14	0	0	13	299	2,856	1,425,963
15	0	0	2	91	941	1,415,162
16	0	0	0	16	333	1,425,536
17	0	0	0	4	108	1,425,480
18	0	0	0	2	44	1,424,266
19	0	0	0	0	8	1,424,392
20	0	0	0	1	6	1,425,044
21	0	0	0	0	1	1,425,327
22	0	0	0	0	0	1,425,568
23	0	0	0	0	1	1,425,068
24	0	0	0	0	0	1,424,803
25	0	0	0	0	0	1,424,248
26	0	0	0	0	0	1,424,738
27	0	0	0	0	0	1,424,738
28	0	0	0	0	0	1,424,730
29	0	0	0	0	0	1,424,089
30	0	0	0	0	0	1,424,313

The values of $L_{n,k}$ for $k = 6$–30 are omitted due to space limitations. These figures cover all two million words of data listed in Table 1.

In this sense, the set of $P_{n,k}$ for a given k defines a probability distribution. On the other hand, Q_n, the frequency of complete words totalling n syllables, does not have this property, and consequently we will refer to Q_n as a (normalized) 'frequency'. Corresponding to Equation 4, we find the following relationship:

$$Q_n = \sum_{k=1}^{n} P_{n,k}. \tag{7}$$

In Fig. 1, we give the plot of $P_{n,k}$ and Q_n for all the data listed in Table 1. The most remarkable global feature of this frequency distribution is its flatness, i.e. its independence from n. Sequences of complete words totalling, say, four syllables are about as common as sequences totalling, say, five syllables, and this is true for all values of n. In order to analyse this structure, we may define an idealized constant distribution $\bar{Q}_n = q$, where $q = 0.720316$ is the average value of the actual distribution Q_n for $n = 1$–30. In the following section, we will examine what lies behind this constant distribution.

2.3 Random-ordering hypothesis

As a working hypothesis, we might assume that the word-length series is randomly ordered, or, more accurately, that 'the number of syllables in a word is independent of the number of syllables in preceding or succeeding words.'

This hypothesis suggests that there is no correlation between the syllable count values in the data series. Such a random-ordering hypothesis allows us to express $P_{n,k}$ in terms of $P_{n,1}$, the probability of a word having n syllables (hereafter we denote this quantity by $p_n = P_{n,1}$). For example, two consecutive one-syllable words satisfy the two-syllable line definition rule. The number of one-syllable words is Ip_1, and according to the random-ordering hypothesis the probability of having a one-syllable

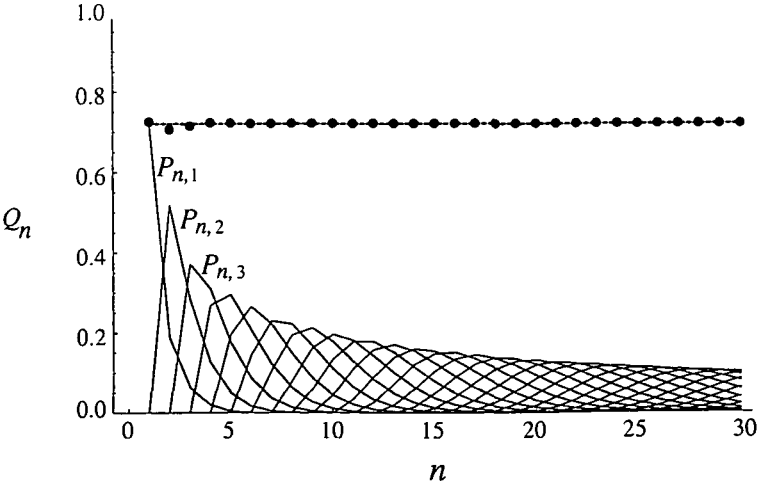

Fig. 1 The frequency distribution for all the data in Table 1. The round dots show the normalized frequency Q_n, the solid lines the probability distributions $P_{n,k}$ for $k = 1, 2, 3, \ldots, 30$. The value of Q_n is obtained by adding $P_{n,k}$ vertically, in other words, adding $P_{n,k}$ for $k = 1,2, \ldots, n$ according to Equation 7. The horizontal dashed line shows the average value q 0.720316 of Q_n.

word after a one-syllable word is not affected by the first word having one-syllable, and therefore is p_1. Thus the number of two-syllable lines of this form is given by $Ip_1 \times p_1$. Dividing this by the total number of words, I, we obtain the normalized frequency $P_{2,2}$:

$$P_{2,2} = {p_1}^2. \tag{8}$$

For larger n and k, combinatoric considerations must be addressed. For example, a three-syllable line can be created by having a two-syllable word and a one-syllable word in sequence, or vice versa. Counting all possibilities, we obtain,

$$P_{3,2} = 2p_1p_2. \tag{9}$$

Some of the other relationships are listed in Table 3. The general expression for $P_{n,k}$ can be obtained in a straightforward manner, but is complicated in written form, and can be handled more efficiently with generating functions.

We define a generating function $P_k(x)$ to represent $P_{n,k}$ for $n = 1-\infty$ by the following:

$$P_k(x) \equiv \sum_{n=1}^{\infty} P_{n,k}\, x^n. \tag{10}$$

Knowledge of all $P_{n,k}$ is equivalent to knowing the behaviour of $P_k(x)$ near the origin $x = 0$, as $P_{n,k}$ can be expressed as the n-th order derivative of $P_k(x)$ at $x = 0$:

$$P_{n,k} = \frac{1}{n!} \frac{d^n P_k}{dx^n}(0). \tag{11}$$

The normalization condition (Equation 6) of $P_{n,k}$ is expressed as $P_k(1) = 1$.

We also define a generating function $Q(x)$ as follows:

$$Q(x) \equiv \sum_{n=1}^{\infty} Q_n x^n. \tag{12}$$

In terms of these generating functions, the relationship in Equation 7 is written as

$$Q(x) \equiv \sum_{k=1}^{\infty} P_k(x). \tag{13}$$

A relationship similar to Equation 11 also holds for Q_n.

Table 3 Some of the consequences of the random-ordering hypothesis

	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
$P_{n,1}$	p_1	p_2	P_3	p_4	p_5
$P_{n,2}$	0	${p_1}^2$	$2p_1p_2$	$2p_1p_3 + {p_2}^2$	$2(p_1p_4 + p_2p_3)$
$P_{n,3}$	0	0	${p_1}^3$	$3{p_1}^2p_2$	$3({p_1}^2p_3 + p_1{p_2}^2)$
$P_{n,4}$	0	0	0	${p_1}^4$	$4({p_1}^3p_2 + 6{p_1}^2{p_2}^2)$
$P_{n,5}$	0	0	0	0	${p_1}^5$

The probability $P_{n,k}$ is listed at the (k, n) position.

The general expression of $P_{n,k}$ in terms of p_n induced by the random-ordering condition can be summarized very simply in terms of the generating functions:

$$P_k(x) = P_1(x)^k. \tag{14}$$

Normalization is trivial in this case: $P_k(1) = P_1(1)^k = 1$. The relationship in Equation 14 leads to the following expression of the generating function $Q(x)$:

$$Q(x) = \sum_{k=1}^{\infty} P_k(x) = \sum_{k=1}^{\infty} P_1(x)^k = \frac{P_1(x)}{1 - P_1(x)}. \tag{15}$$

Thus, the reason for introducing the random-ordering hypothesis becomes evident. If that hypothesis is valid, the relationship in Equation 15 can be used to explain the features of the frequency distribution Q_n in terms of the features of the one-word probability distribution p_n.

We will now turn to the verification of the random-ordering hypothesis. In Table 4 we list the number of n-syllable words following immediately after m-syllable words, i.e. the number of one-syllable words after one-syllable words, two-syllable words after one-syllable words, and so on; and then the number of one-syllable words after two-syllable words, and so on. The corresponding probability distribution is plotted in Fig. 2. From this figure, it may be readily observed that the distribution is close to being independent of the value of m, and we therefore conclude that the random-ordering hypothesis is valid to a reasonable degree of accuracy.

The reader should note that these features, the flatness of the frequency Q_n and the random ordering, are also true for individual authors and works, and are not a result of averaging over them. In Figs 3 and 4, we give the plots of Q_n and $p_{n,m}$, respectively, for George Eliot's *Middlemarch*, where the relevant features are clearly apparent, as they are for all the texts in our corpus.

Table 4 List of the number of occurrences of n-syllable words after m-syllable words

m	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$
All	1,433,426	371,500	122,179	40,314	9,048	1,082	119	6	2
1	1,025,719	279,915	90,473	29,731	6,680	811	91	5	1
2	278,764	63,357	20,960	6,733	1,485	183	17	0	1
3	92,302	19,542	7,263	2,422	589	52	8	1	0
4	29,414	6,815	2,710	1,128	217	27	3	0	0
5	6,400	1,617	692	265	67	7	0	0	0
6	745	225	71	30	9	2	0	0	0
7	75	28	10	5	1	0	0	0	0
8	5	1	0	0	0	0	0	0	0
9	2	0	0	0	0	0	0	0	0

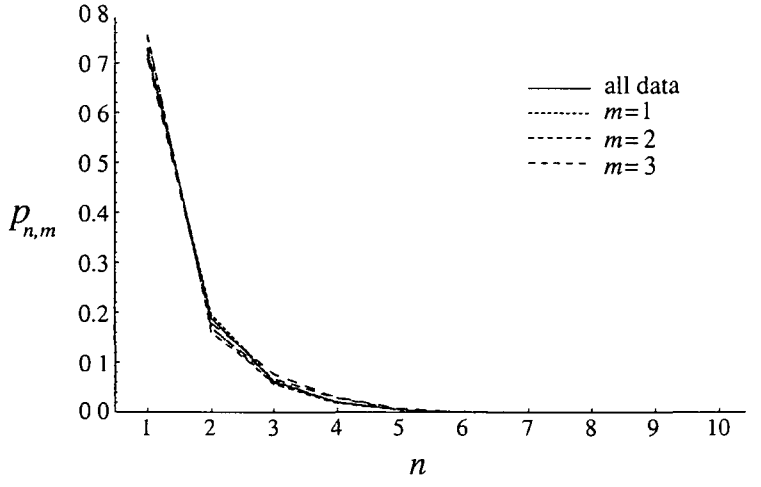

Fig. 2 Plot of the probability distribution $p_{n,m}$ for the data in Table 4. The solid line shows p_n, while other lines show $p_{n,m}$.

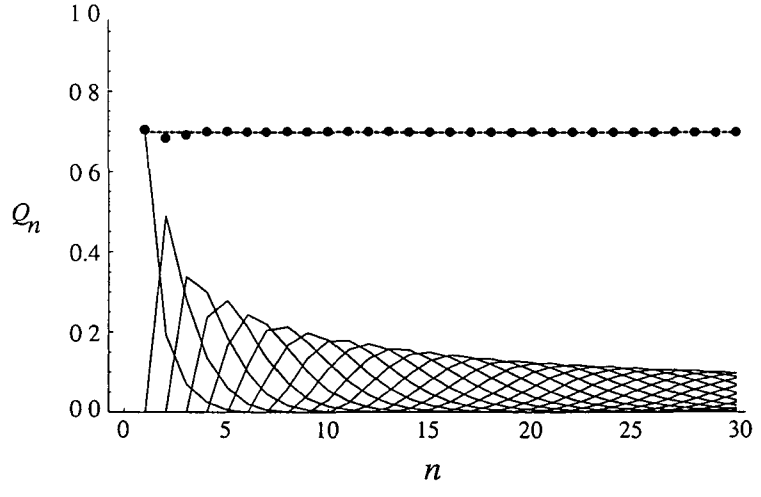

Fig. 3 Plot of the normalized frequency Q_n and the probability distributions $P_{n,k}$ for George Eliot's *Middlemarch*. The horizontal dashed line shows the average value $q = 0.69844$ for Q_n.

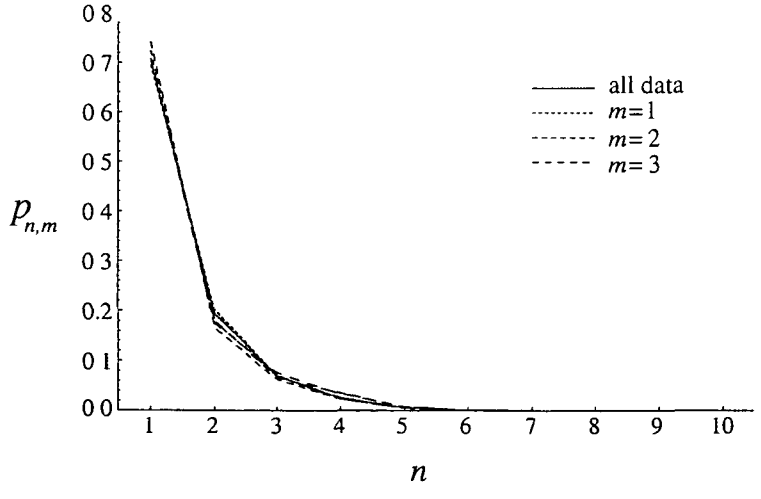

Fig. 4 Plot of the probability distribution $p_{n,m}$ for George Eliot's *Middlemarch*.

2.4 Single word probability

Now that the random-ordering hypothesis has been confirmed, we can obtain the probability $\bar{p}_n$ that induces the constant frequency distribution $\bar{Q}_n = q$. The generating function for $\bar{Q}_n$ is as follows:

$$\bar{Q}(x) = \sum_{n=1}^{\infty} qx^n = \frac{qx}{1-x}. \tag{16}$$

By solving Equation 15 in terms of $P_1(x)$, we obtain,

$$\bar{p}_1(x) = \frac{\bar{Q}(x)}{1+\bar{Q}(x)} = \frac{qx}{1-(1-q)x} = \sum_{n=1}^{\infty} q(1-q)^{n-1}x^n. \tag{17}$$

Therefore,

$$\bar{p}_n = q(1-q)^n - 1, \tag{18}$$

which is the geometric probability distribution. In Fig. 5, we compare the actual probability distribution p_n (dots) and the theoretical geometric distribution $\bar{p}_n$ given by Equation 18 (dash-dotted line), which are in close agreement. Further, the geometric distribution (Equation 18) yields the average number of syllables per word (mean word length) as follows:

$$\langle n \rangle = \sum_{n=1}^{\infty} np_n = \bar{P}_1'(1) = \frac{1}{q}. \tag{19}$$

This relationship was noticed earlier by Constable (1997, p. 182), when he stated that the number of complete words totalling n syllables in English prose was given by dividing the total number of words by the mean number of syllables per word. Accounting for this is one of the goals of the present paper, and by adopting the random-ordering hypothesis we arrive at a new finding, namely the relationship between the constant distribution $\bar{Q}_n$ and the geometric distribution $\bar{p}_n$, which gives a sound theoretical basis to the earlier observation.

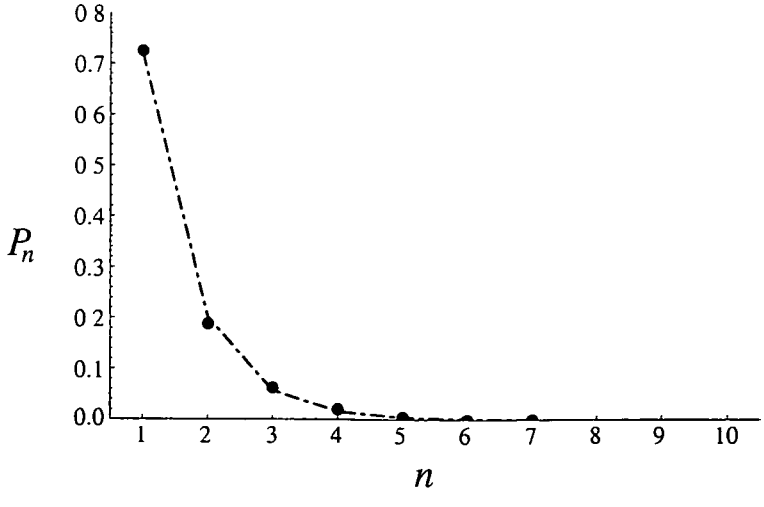

Fig. 5 The probability distribution $\bar{p}_n$ (denoted by dots) and the theoretical prediction $\bar{p}_n$ (Equation 18) (denoted by the line).

2.5 Interpretation: random segmentation

The two global properties described above, the random-ordering and the geometric distribution (Equation 18), allow a definite characterization of the word length data, since these are the properties typical of a system with a given probability of termination at any point: namely, if one assumes that sequences of syllables are constructed such that after any syllable the end of a word happens with probability q, the above geometric distribution (Equation 18) is obtained. Putting this in a slightly different manner, if one has a large number of syllables and word boundaries (spaces) in a $(1 - q)$ to q ratio and randomly places them in sequence, the same distribution is obtained. More straightforwardly still, the segmentation of English prose into words is a random phenomenon. This is not to suggest that the segmentation of English is in all respects fundamentally random; our hypothesis merely notes that whatever principle of regular order may be operating elsewhere, perhaps in relation to stress or phonemes, word boundaries and syllable boundaries are related with a fixed probability.

2.6 Word length frequencies in output and in the lexicon

The fact that the observed distribution of word length frequencies in texts is not found in the lexicon itself, which is plotted in Fig. 6, has been noted by other researchers (Wimmer *et al.* 1996), and provoked explanation in terms of attractors and control cycles in the composition process. Strictly speaking, it is beyond the scope of our paper to engage deeply with this question, but since these researchers have mistakenly presumed that the output frequency distribution is itself ordered, it seems worthwhile to point out that hypotheses explaining the relationships between the output and lexicon distributions in terms of the operation of randomness deserve, at the least, serious consideration.

For example, we might speculate that the concept of 'word' is a relatively late (though perhaps prehistoric) analytical category, and was

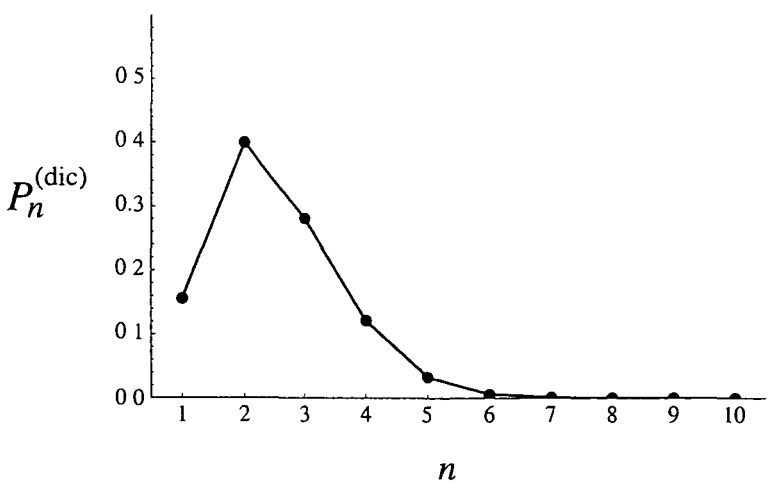

Fig. 6 Probability distribution for words with n-syllables in Constable's lexicon

arrived at by segmenting the verbal output stream in such a way that word boundaries were placed with a fixed probability in relation to syllable boundaries. The resulting word forms were accumulated to compose a lexicon. Since the sound system of a language is not infinitely extendible, there were more unique and acceptable disyllabic forms than monosyllables, more trisyllables than disyllables, and so on. Thus, although in its early stages, the lexicon would, obviously, have followed the geometric distribution of the output from which it was drawn (Fig. 5), eventually it would adopt a curve resembling that found in our data (Fig. 6).

However, most of the more commonly occurring language functions will have been segmented into monosyllabic words, and thus, assuming that modern speakers and writers are employing more or less the same language, and using language for the same functions, the proportion of their output which is monosyllabic will remain steady. In such a view, the stability of the geometric distribution in modern output is an artefact of the process of random segmentation and accumulation which resulted in the lexicon, and does not result from a bias in the decision process by which words are selected from the lexicon. Indeed, from this perspective, it would seem that speakers are not, in the strict sense, selecting 'words' from the lexicon, but are choosing other linguistic objects which happen to have been segmented with a fixed probability.

Of course, the process of accumulation may be complicated by the adoption of words from other languages and also by the fact that words may fall into disuse or be transformed in ways which are non-random with regard to length. Consequently, no straightforward match between the model outlined above and our data is to be expected. Nevertheless, preliminary examination suggests that even with a fairly simple set of assumptions, a promising global approximation to our lexicon can be obtained.

Imagine the set of all the discourse produced in a language to date, and denote the number of words in this canonical set by H. Secondly, assume that there are k monosyllables, and that any combination of these monosyllables will make acceptable words. Thirdly, assume that the words are terminated with a probability q. These three assumptions allow us to compute the expected number of words of any given length. For example, the expected number of monosyllables is $qH\ (\equiv H_1)$. The probability of all the words of one length being identical is $k \times (1/k)^{H_1}$, the probability that the monosyllables comprise repetitions of two word forms is ${}_kC_2 \times [(2/k)^{H_1} - (2(1/k)^{H_1}]$, and so forth. A straightforward calculation shows that the expected number of different monosyllabic words in the lexicon is then given by the following;

$$k\left[1 - \left(1 - \frac{1}{k}\right)^{H_1}\right] \simeq k\left[1 - e^{-H_1/k}\right], \tag{20}$$

where the last approximation is valid for $k \gg 1$, which applies to the following analysis. Similarly, for n-syllable words, k and H_1 are replaced

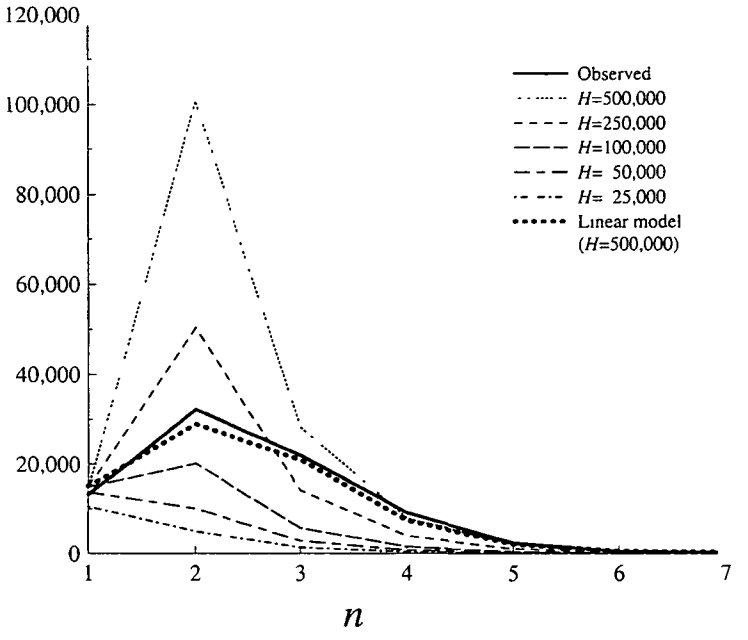

Fig. 7 The actual lexicon distribution (solid line) and some results from our model, including the linear model (heavy dotted line).

by $k_n = k^n$ and $H_n = q(1 - q)^n H$, respectively. This yields the lexicon distribution with three free parameters q, k, and H. Some of the results generated by the model are plotted in Fig. 7. In this plot, we have used the value $q = 0.72$ derived from our corpus data, and the value of k has been set at 15,000, the value required, approximately, to explain the number of monosyllables in our lexicon. Plots are given for canonical sets H which range from 25,000 to 500,000 words. These results are plotted as normalized figures in Fig. 8, enabling comparison with the theoretical geometric distribution and the normalized lexicon distribution. As can be seen, low values for H produce a good approximation to the geometric distribution, while higher values produce theoretical lexicons resembling the overall structure of that observed in our data, while differing significantly in their quantities. Closer approximations can be achieved with simple and plausible alterations to the model. For example, the numbers of short words, monosyllables (and, for some values of H, disyllables) appearing in the lexicon are essentially given by k_n, as can be seen from Equation 20 with the approximation $k_n \ll H_n$. Therefore, this part of the lexicon distribution depends heavily on the modelling of the construction of polysyllables from monosyllables. Since not every combination of monosyllabic words will produce allowable well-formed polysyllables, one may reasonably modify the second assumption and employ a function of n which increases less rapidly than the power function k^n. For example, we could use a linear function nk, which assumes simply that allowable disyllables are exactly twice as frequent as monosyllables, and in fact this generates results in close agreement with the actual lexicon derived from our data (see the heavy dotted line in Fig. 7).

We recognize that this hypothesis is strongly counter-intuitive in

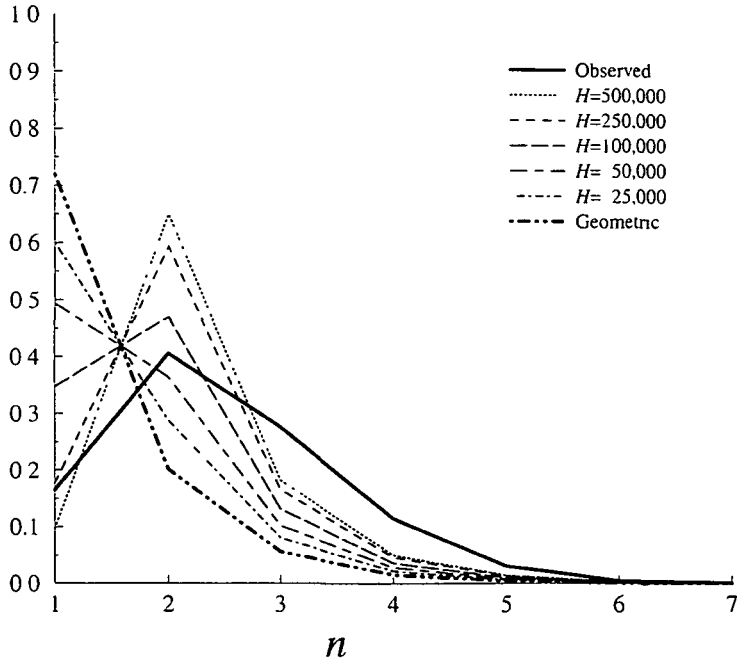

Fig. 8 The normalized plot of the observed lexicon distribution (solid line) and some results from our model, together with the theoretical geometric distribution with $q = 0.72$.

suggesting that word boundaries as defined by spaces do not segment the language at linguistic joints, but we believe that the economy of its explanatory power, even in this rudimentary form, is strongly in favour of hypotheses of this type, and justifies further investigation along these lines.

3 Fine Structures

Readers will have noticed the differences between the global structure of Q_n and the actual distributions, the most notable of which is the small dip at Q_2 below the average value q seen in Fig. 1, and the small differences between the $p_{n,m}$ of different values of m in Fig. 2. In plain terms, sequences of words totalling two syllables appear to be somewhat less common than other totals, and thus there is evidence that the length of a word has some effect on the value of preceding and succeeding words.

We need first to guard against the possibility of being misled by statistical errors, and determine whether these differences are meaningful quantities requiring causal explanation, or can simply be attributed to statistical fluctuations. In the following, we first discuss the handling of statistical errors and then proceed to the discussion of features of the fine structure.

3.1 Statistical errors for Q_n

The standard estimate for statistical errors may be applied to the probabilities $P_{n,k}$, and presents no problems, but the estimate of the error range for the frequencies requires further discussion. The relationship in Equation 7 enables this procedure, and is trivial for Q_1, as it is actually a

probability, $Q_1 = P_{1,1}$. Therefore, the standard deviation for Q_1 is given by $\sigma = \sqrt{Q_1(1 - Q_1)} / \sqrt{I}$.

According to Equation 7 that for $Q_2 = P_{2,1} + P_{2,2}$. The probability $P_{2,1}$ is given by dividing the observed number $L_{2,1}$ of two-syllable words by the total number of words I. From Fig. 5, we see that $P_{2,1}$ is ~0.2; therefore, the standard deviation of $L_{2,1}$ can be approximated as $\sigma_{2,1} \sim \sqrt{L_{2,1}}$. Similarly, we can state approximately that $\sigma_{2,2} \sim\sqrt{L_{2,2}}$. Thus the total standard deviation σ_2 for the observed number of strings $L_2 = L_{2,1} + L_{2,2}$ is given by $\sigma_2^2 = \langle L_2^2 \rangle - \langle L_2 \rangle^2 = L_{2,1} + L_{2,2} = L_2$, and calls on the statistical independence of $L_{2,1}$ and $L_{2,2}$, $\langle L_{2,1} L_{2,2} \rangle = \langle L_{2,1} \rangle \langle L_{2,2} \rangle$. Consequently, the standard deviation of Q_2 is given simply by $\sqrt{Q_2 / I}$, just as if it were a probability. The same is true for other values of Q_n.

3.2 Deviations from the flat Q_n distribution

To clarify the small deviation of the frequency Q_n from the flat distribution, we can plot the difference between the actual frequency and the theoretical flat distribution, $\delta Q_n = Q_n - q$, with vertical bars showing the 3σ error ranges, as in Fig. 9. From this figure, we find that the Q_2 depression is statistically significant, as are other deviations at $n = 1$, 3, and 4. The generality of these deviations may appear to be doubtful, and there are certainly a number of unanswered questions. The corpus is predominantly of high status literary writing, mostly nineteenth century, and of that a substantial portion comes from one author, Henry James. It might therefore be suggested that some of these deviations are characteristic of an output type, or a period, or even of an author. However, the Q_2 depression is found consistently across the authors and works we have examined, though in the case of Kipling we found a significant depression at $n = 1$ instead of an enhancement. We predict, therefore, that the Q_2 depression is a universal characteristic, or nearly so. It seems possible,

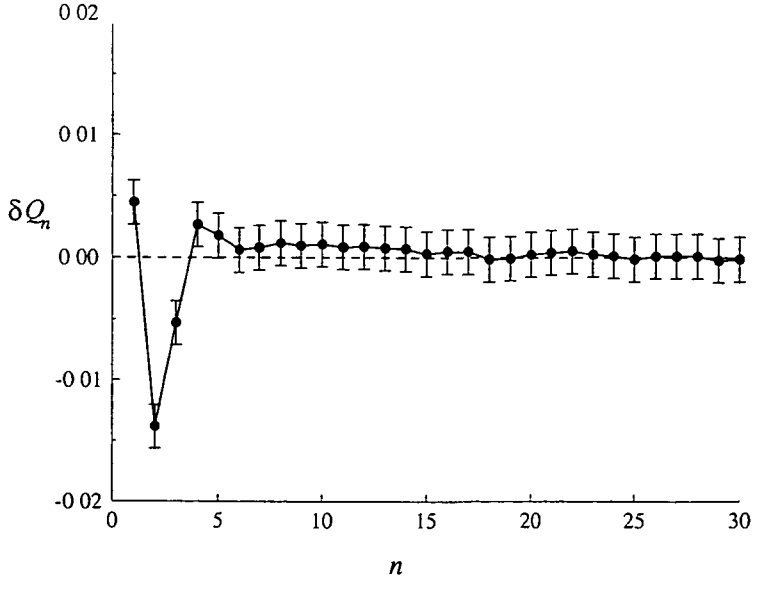

Fig. 9 Detail of $\delta Q_n = Q_n - q$. The vertical bars show the 3σ-confidence ranges of statistical errors for each data point. Note that the horizontal range covers only one twenty-fifth of the range of Fig. 1.

indeed likely, that some of the other deviations will prove to be particular to an author, a work, a genre type, or a period, but their universality should not be ruled out at this stage.

While a refined discussion of the causes underlying the Q_n deviations observed in our data must await the decision as to which are universal, a general explanation can be offered in terms of the underlying deviations, namely: (i) deviation from the geometric distribution; and (ii) deviation from random ordering. To bring these issues into focus, Fig. 10 plots (a) $p_n - \bar{p}_n$ (solid line) and (b) $p_{n,1} - \bar{p}_n$ (dotted line). Line (a) is a measure of the deviation from the geometric distribution, while the difference between (a) and (b) is a measure of the deviation from random ordering. In these figures, we find that (i) monosyllabic words have a slightly higher probability than that predicted by the theoretical geometric distribution, while disyllabic words have a slightly lower probability, and (ii) in the sequential distribution there is a slightly enhanced probability of a polysyllable after a monosyllable (relative to random ordering). It is straightforward to prove that these small perturbations do not affect Q_n for large values of n, a fact which is hardly surprising since such values reflect long-range correlations between word lengths, and our results show that correlations of this kind are almost completely absent.

The interplay of the two causal factors, (i) and (ii) above, deserves comment, if only to prepare the ground for further discussion elsewhere, and can be studied as follows. Taking our *Middlemarch* data, the largest single text of those we have studied, we find that it has $Q_1 = p_1 = 0.7043$, and $Q_2 = 0.6832$; the difference being $\Delta Q_2 = Q_1 - Q_2 = 0.0210$. We may begin by examining the effect of cause (i), the surplus of monosyllabic words, by employing the random-ordering hypothesis but taking into

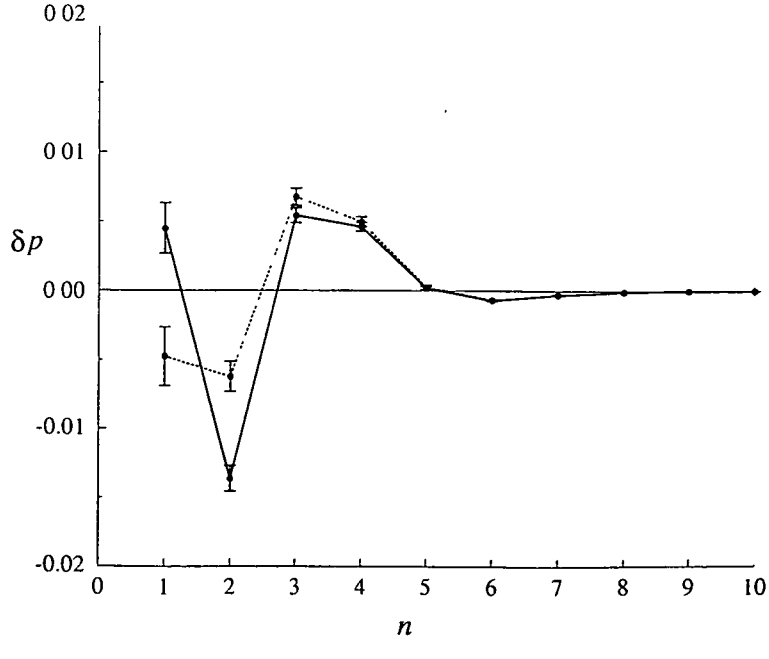

Fig. 10 Detail of (a) $p_n - \bar{p}_n$ (solid line) and (b) $p_{n,1} - \bar{p}_n$ (dotted line).

account the deviation of p_n from the ideal geometric distribution. According to the random ordering hypothesis, we have $Q_2 = p_1^2 + p_2$. Using the observed value $p_2 = 0.1938$, we find that $Q_2 = 0.6898$. While this value explains 0.0145 of ΔQ_2, it is not sufficient. We may now take into account cause (ii), the deviation from random ordering, by employing the formula for Q_2 which is $Q_2 = p_1 p_{1,1} + p_2$. Using the observed value $p_{1,1} = 0.6950$, we find that this yields $Q_2 = 0.6832$, which agrees with the observed value.

We have not conducted a thorough examination of the underlying linguistic causes of these deviations, but candidates spring readily to mind and consequently some cautionary remarks are in order. With regard to the deviation from random sequencing, it might appear that the relationship between commonly occurring function terms, which are predominantly monosyllabic, and content terms, which are somewhat more likely to be polysyllabic, is a probable explanation, but the situation is by no means straightforward, as the following consideration will make clear.

Suppose that one hypothesizes that cause (ii) is due to correlations between articles and their subsequent words. We find that in the *Middlemarch* data, after the articles 'the', 'a', and 'an', the p_n distribution is (0.4072, 0.3521, 0.1645, 0.0598, 0.0149, 0.0013, 0.0002), which certainly shows a significant shift towards polysyllables, as compared with the overall random distribution. If we assume that the word length distribution after other monosyllabic words has the normal p_n distribution, we can calculate the induced $p_{1,1}$ by $p_{1,1} = \alpha p_{1,1}^{\text{article}} + (1 - \alpha)p_1$, where α is the ratio of the number of articles to the total of monosyllabic words, and $p_{1,1}^{\text{article}} = 0.4072$ is the probability of monosyllabic words after articles. Using the observed value $\alpha = 0.1028$, we find that $p_{1,1} = 0.6738$. This suppression of $p_{1,1}$ compared with p_1 exceeds the actual depression observed, and thus we may conclude that there are other significant structures affecting the Q_2 depression and cancelling it. The nature of these other structures is at present unknown.

Whatever the best causal account, it should be emphasized that deviations such as these, and others of the same kind which might be discovered by future research, are subtle variations from a strong fundamental trend, that of randomness, and it is not safe to conclude that they are evidence which 'confirm[s] the assumption of a non-accidental distribution of word lengths' (Ziegler, 1996, p. 73).

3.3 Test for lineation

Apart from the $n = 2$ deviation observable in the prose texts in our corpus, we are aware of one large class of texts which routinely exhibit significant deviations in the Q_n distribution, namely isometrically lineated verse texts. From our perspective, this is hardly surprising. Texts composed in regular lines are by definition ordered with respect to lineation, and provided that the text is composed in lines of ten syllables, for example, then $n = 10$, and all multiples of ten, will be substantially above the flat distribution, and if it is composed in two core line lengths, as are limericks and Spenserian stanzas, then it will exhibit two series of peaks.

It may be noted, however, that the existence of these peaks in verse is of theoretical rather than a practical value. It is unlikely that we will often wish to test in order to detect lineation, since lines are usually visually evident, or revealed by other features such as rhythm. However, as a contribution to the theoretical definition of lineated as distinguished from unlineated texts (verse as distinguished from prose), the procedure is of considerable interest. Although it has long been obvious that lineation is not merely 'a visual or typographical fact' but a 'fact of the language' (Wimsatt and Beardsley, 1959, p. 591), to use one well-known formula, there has been, to our knowledge, no conclusive empirical demonstration of the presence of this fact, or any theoretical explanation of its character. The deviation from the flat distribution performs both these functions, and will be discussed at greater length in Part II of this paper, where we will present examples and offer a detailed analysis of the dynamics underlying peak creation.

4 Conclusion and Comments

Previous research on word length distribution (Wimmer and Altmann, 1994; Becker, 1996; Best and Altmann, 1996; Dittrich, 1996; Frischen, 1996; Riedemann, 1996; Rottger, 1996; Wimmer *et al.*, 1996; Ziegler, 1996; Zuse, 1996) has attempted to infer significance from what we suggest is an improperly identified non-geometric curve, and held that this curve supports the belief that 'language is [. . .] a self-regulating system, which is controlled by the needs of the language community' (Zuse, 1996), or is an organism of interrelated control cycles (Wimmer *et al.*, 1996). Furthermore, these researchers have borrowed terms from chaos theory in ways that are seriously misleading. For example, in Wimmer *et al.* (1996, p. 98), it is claimed, on the basis of a small data set, that 'the sequence of words is clearly chaotic', and that the distribution of word length in a text could be explained by reference to 'attractors'. However, as we have made clear, there is in fact no chaos in its mathematical sense and what we observe in our study is randomness. We cannot rule out the discovery of the sort of order sought by the synergetic linguists, but must observe that our findings give little support to the supposition of its existence in relation to word length. Thus, in approaching language frequency data of this type, we find ourselves generally in sympathy with those such as Mandelbrot (1961) and Li (1992) in their advocation of interpretations grounded in the operation of random factors, and we are less drawn to positions such as those proposed by Zipf (1965, p. 48), where statistical regularities are seen to arise from some deep principle of order.

In conclusion, however, we should like to emphasize that the theory and data outlined here are of more than negative value. Our investigation was derived from empirical observations and hypotheses offered in an earlier paper (Constable, 1997) with regard to the process of making verse lines, and those remarks are confirmed by our results. The relationship between the mean number of syllables per word and the number of

sequences of words totalling a given number of syllables (Constable, 1997, p. 182) is dependent on the geometric frequency of word length totals and the random distribution of those words in the text sequence, which we have shown here to be solid findings. Thus, the apparently arcane facts of word length distribution in English output can be seen to deepen our understanding of one, and historically a very important, area of language output, isometrical verse. Finally, and perhaps most interestingly for literary scholars, the establishment of the randomness of the distribution of spaces in English permits the characterization of a property of prose against which verse may be distinguished.

References

Attridge, D. (1982). *The Rhythms of English Poetry*. London: Longman.

Attridge, D. (1995). *Poetic Rhythm: An Introduction*. Cambridge: Cambridge University Press

Becker, C. (1996) Word lengths in the letters of the Chilean author Gabriela Mistral. *Journal of Quantitative Linguistics*, 3: 128–31.

Best, H.-H. and Altmann, G. (1996). Project report. *Journal of Quantitative Linguistics*, 3: 85–8.

Constable, J. (1997). Verse form; a pilot study in the epidemiology of representations. *Human Nature*, 8: 171–203.

Dittrich, H. (1996). Word length frequency in the letters of G. E. Lessing. *Journal of Quantitative Linguistics*, 3: 261–4.

Frischen, J. (1996). Word length analysis of Jane Austen's letters. *Journal of Quantitative Linguistics*, 3· 80–4

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38: 1842–5.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. In: *Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics*, American Mathematical Society, 12: 190–219.

Riedemann, H. (1996). Word length distribution in English press texts. *Journal of Quantitative Linguistics*, 3· 265–71.

Röttger, W. (1996). Distribution of word length in Ciceronian letters. *Journal of Quantitative Linguistics*, 3: 68–72.

Wimmer, G. and Altmann, G. (1994). The theory of word length: some results and generalizations. *Glottometrika*, 15.

Wimmer, G., Kohler, R., Grotjahn, R., and Altmann, G. (1996). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1: 98–106.

Wimsatt, W. K. and Beardsley, M. C. (1959). The concept of meter: an exercise in abstraction. *PMLA*, 74: 585–98.

Ziegler, A. (1996). Word length distribution in Brazilian–Portuguese texts. *Journal of Quantitative Linguistics*, 3: 73–9.

Zipf, G. K. (1965). *The Psychobiology of Language*. Cambridge, MA: MIT Press.

Zuse, M. (1996). Distribution of word length in early modern English: letters of Sir Philip Sidney. *Journal of Quantitative Linguistics*, 3: 272–6.