WORDS AND SYLLABLES : FOUR ECLOGUES

The major result of automatic scansion is the production of a completely scanned text in machine-readable form, which is then available for further analysis by means of the computer. What follows here is a pilot study (in which the computer was not in fact used) to illustrate one direction in which further analysis may proceed. I have chosen to examine here what seems to me to be a feature of Latin hexameter poetry which is readily amenable to statistical analysis, despite its, initially at least, complete lack of intrinsic interest. As everyone knows, the hexameter line consists of from 13 to 17 syllables (barring certain exceptional cases), and it usually requires about 6 or 7 words to fill the line. This means that the average Latin word in hexameter poetry consists of two and a fraction syllables (and, for all I know, this may be true of Latin words in general). So far as words and syllables go, this seems to be about as a much as one needs to know, and in fact it is not at all clear that one needs to know this much in order to appreciate Latin poetry in any reasonable way. After all, so much else of importance is happening in the poetry that it seems absurd to spend much energy in observing this particular feature. Indeed, it seems reasonable to assume, at least as an initial hypothesis, that the authors

themselves did not pay much attention to this aspect of their word. So much then as an introduction meant both to disarm and warn the reader, In what follows, I shall pay a great deal of attention to the relation between words and syllables in Eclogues 1, 4, 6 and 9 (a largely arbitrary selection). The justification, if any, will emerge only in the course of the investigation.

First, a few ground rules. The text is the Teubner edition. A word is what is printed as such in that edition, i.e., a group of letters preceded and followed by a space or mark of punctuation. When elision or ecthlipsis occurs, the preceding word loses a syllable. Prodelision is not recognized; it is taken as simple elision.

Here are some of the basic facts in summary.

TABLE 1

	Eclogue 1	Eclogue 4	Eclogue 6	Eclogue 9	Total
Lines	83	63	86	67	299
Words	559	409	558	472	1998
Syllables	1235	924	1276	1016	4451

From these one quickly derives the following averages or means.

TABLE 2

	Eclogue 1	Eclogue 4	Eclogue 6	Eclogue 9	Total
Words per line	6.73	6.49	6.49	7.04	6.68
Syllables per line	14.88	14.67	14.84	15.16	14.89
Syllables per word	2.21	2.27	2.29	2.15	2.23

Let us pause to consider these figures. Comparison shows that Eclogue 1 adheres most closely to the figures derived for all four poems combined, and it is at least noteworthy that the poem which heads the collection has this quality. I do not mean to imply any farfetched mystical or numerological conclusion here. In fact, a rapid check reveals that the average number of words per line for the entire work is 6.79 and that Eclogue 10 (with 6.77) comes closer, while Eclogue 2 also has 6.73 words per line. (I have not counted all the syllables of all the Eclogues.) There is no implication here that Vergil counted his syllables or that he wrote with an abacus constantly at his elbow. Nevertheless, there is an implication of a sort here and it should be clarified at this point. Despite the introduction, I do not feel that the relation between syllables and words is utterly without point, nor is it completely bereft of critical utility. The operative word here is "critical". The application of statistical procedures to literature is not new, particularly with regard to classical literature, but, ordinarily, resource to statistics is made in order to buttress arguments

concerning chronology or disputed authorship, questions which are important but, in a way, only ancillary to the process of criticism itself. I shall not be concerned with problems of chronology here, and I have chosen four poems by a single author so as to avoid the issue of authorship. Instead, I am concerned here with one facet of a critical entity which may be called texture. It is a rather mysterious entity. As St. Augustine remarked on the concept of time, everyone knows what it is until asked to define it. Clearly, texture has something to do with pacing or tempo, and we have all had impressions of comparative rapidity or stateliness when exposed to a work of art (it need not be great art). The notions of complexity and simplicity are also clearly involved, and perhaps above all, the Peripatetic notion of *τὸ πρέπον*, of a concordance or harmony of ideas or conceits and the formal means by which they are expressed. This is to raise a problem of huge critical and theoretical difficulty which is further complicated by some of its relativistic aspects, for, after all, one era's appropriate form may be viewed by another as intolerably cluttered or excessively bald. To make my point very quickly and summarily, if it is the role of criticism to attempt to describe and analyze these qualities, then it seems possible that some of these descriptions and analyses may most appropriately be given in statistical form. To be sure, statistical terminology is not the language we customarily use in discussing literary issues, but this does not seem to be a serious objection.

The implication may now be stated more clearly. Someone (presumably Vergil) placed *Eclogue 1* at the beginning of the collection because (again presumably) he thought this to be an appropriate position for that poem. We do not know why, although it is the business of criticism to try to

make intelligent guesses. It is difficult to believe (and I do not believe) that this position is due to the fact that the mean number of words per line in the poem is 6.73. But that number does represent a small fact about the texture of the work and it is not inconceivable that this fact, almost surely not consciously perceived, contributed its mite to that larger aesthetic decision which placed the poem first in the collection. In any case, as said above, it is noteworthy that this poem chosen in some sense to represent the entire work happens to possess a statistical characteristic which also tends to represent the entire work. So much as a first tentative annotation.

The extremes in Table 2 are also of interest. Every entry for *Eclogue 9* is an extreme. It has the shortest average word-length (measured in syllables); and the longest average line-length (whether measured in number of syllables or in number of words). At first sight, however this last mode of measurement seems absurd. Since the average word-length differs, it is as though we were to measure something using sometimes a long inch and sometimes a short one, and one has always felt the injustice of the telegraph company charging as much for a short word as for a long one. Hence it should come as no surprise that the average line in *Eclogue 9*, measured as it is in terms of the smallest average word, turns out to be the longest. It is a happy coïncidence however that if we measure in terms of syllables, here too, *Eclogue 9* has the largest average line-length. Unhappily, this pleasing symmetry does not hold when we look at the opposite situation. Although *Eclogue 6* has the largest average word, its average line-length measured in words is equalled in its shortness by that of *Eclogue 4*; and if we measure in syllables *Eclogue 4* has the shortest average line.

It seems fair to conclude that syllables constitute a more legitimate measure than words; and we have tacitly assumed that it is all right to state average word-length in term of syllables. However, Latin verse consists of long and short syllables, and this should be considered.

The ancient equivalence of one long syllable to two short ones is worth consideration, because it makes all hexameter verses of equal duration, a notion which seems intuitively attractive. On this basis, every verse would consist of 23 *morae* if one takes the final syllable as short (which I always do). I have not chosen to pursue that mode of analysis here, but it certainly deserves further discussion. Using *Eclogue 4* as an example, compare the distribution of words according to the number of syllables they contain, and according to the number of *morae* they contain.

TABLE 3

Syllables	No. of words
1	82
2	171
3	126
4	28
5	2
	409

TABLE 4

Morae	No. of words
1	10
2	96
3	94
4	136
5	21
6	48
7	4
	409

The distribution for *morae* in Table 4 has disturbing features in that the figures for 3- and 5-*mora* words do not lie between the figures above and below them. The reason for these disturbances lies in the conventional character of the hexameter verse.

Briefly the hexameter verse does not allow the substitution of two short syllables for a long in the first position of the foot. Hence, short syllables must always occur only in pairs (except for syllaba anceps), and one can never have an even number of long syllables bracketed by short syllables. If this conventional restraint were not in force, the number of possible word forms could be calculated as follows : there is only one possible form for a one-mora word; a two-mora word can occur as a single long or as two short syllables, and so on. In fact, the number of possible forms generated in this way is represented by a Fibonacci series (where each term is the sum of the two preceding ones). However only a fraction of the possible forms are acceptable into the hexameter meter. Here again in tabular form :

TABLE 5

Number of morae in word	Possible form	Acceptable form
1	1	1
2	2	2
3	3	2
4	5	4
5	8	4
6	13	5
7	21	4

That is, while the number of possible forms increases continually, the number of acceptable forms does not, and it should be noted that the disturbances in the distribution of acceptable forms also occur for 3- and 5-mora words. Now, while this is a fairly dramatic representation of the strictures imposed by the hexameter form, in practice these strictures were not so serious, since the poets could always rely on the rules of position to enable them to incorporate all but the most recalcitrant words. It would, therefore be instructive to inspect the ways in which poets have utilized the rules for making position. Certainly it would be of interest to know whether poets or poems differ markedly in their degree of reliance upon one convention in order to overcome the trammels of another but I have not done that here. (The obvious practical difficulty lies in programming the computer to distinguish between natural and positional quantity.)

For the moment, then, we shall use syllables and words as our units, and the *mora* as a unit will be ignored. It is clear that many of the techniques to be discussed below could equally well be applied to a count based upon *morae*.

Returning to Table 2, we note once again that on the average, *Eclogue 9* contains the shortest words and the longest lines (in syllables or in words), that *Eclogue 4* contains the shortest line (in syllables) and that *Eclogue 6* contains the longest average-word (in syllables). Standard statistical description, however, is not confined to the average. Two men may be each 6 feet in height. Another pair may have heights of 5 and 7 feet respectively. In both cases, the average height is 6 feet, but the average

alone is clearly not a satisfactory description. We shall learn something more about the "texture" of the poems in question, if we are given some idea of the degree of dispersion about the various averages given in Table 2. In statistical practice, the unit for describing such dispersion is the standard deviation and some notion of its meaning is helpful (it is discussed in any elementary text book on statistics). By definition, the standard deviation is the positive square root of the average of the squares of the observed deviations from the mean. As an example we display the calculation for the number of words per line in *Eclogue 4.*

TABLE 6

Eclogue 4

Mean words per line = 6.49

(1) No. of words per line	No. of lines observed	Deviation from mean	Square of col. 3	Col. 2 x Col. 4
5	11	-1.49	2.22	24.42
6	24	-0.49	0.24	5.76
7	17	0.51	0.26	4.42
8	8	1.51	2.28	18.24
9	3	2.51	6.30	18.90
	63			71.74

$$\text{Standard deviation} = \sqrt{\frac{71.74}{63}} = 1.06$$

Avoiding all technicalities, a simple interpretation would be that about 2/3 of the time, the number of words found in a line will be within the limits 6.49 + 1.06 and 6.49 - 1.06. This number in itself is not very enlightening, but a comparison for the four Eclogues will help.

TABLE 7

Words per line

	Mean	Standard deviation
Eclogue 1	6.73	0.92
Eclogue 4	6.49	1.06
Eclogue 6	6.49	1.04
Eclogue 9	7.04	1.13

The differences are not great, but for what they are worth, the number of words per line varies most widely in *Eclogue 9*, which is perhaps not surprising, given the greater length of its average line. It is worth noting however, that the number of words per line varies least in *Eclogue 1*, despite its intermediate average line-length.

The standard deviations for the other means in Table 2 are listed below.

TABLE 8

Syllables per line

	Mean	Standard deviation
Eclogue 1	14.88	0.83
Eclogue 4	14.61	0.64
Eclogue 6	14.84	0.91
Eclogue 9	15.16	0.87

TABLE 9

Syllables per word

	Mean	Standard deviation
Eclogue 1	2.21	0.90
Eclogue 4	2.27	0.89
Eclogue 6	2.29	0.93
Eclogue 9	2.15	0.90

Here too it should be remarked that the number of syllables per line varies least in *Eclogue 4* and most in *Eclogue 6*. The same is also true for the variation in the number of syllables per word. These figures, then, allege a greater degree of uniformity of texture in *Eclogues 1* and *4* than in *Eclogues 6* and *9*. It must be left to the reader to decide whether this

finding coïncides with the impression received in a subjective reading of the poems.

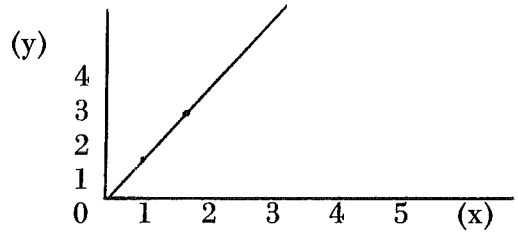
We now turn to a question which will hopefully throw light on the discussion to follow. Are long lines (measured in syllables) longer because they contain more words or because they contain longer words, or both ? If both, which factor is more important : the length of the words or the number of the words ? At first glance, I should think that this question does not even seem to make sense. But a moment's reflection will show that the question is not completely senseless, although its utility may indeed be dubious. With regard to the last point, however, we shall end with a set of indices which serves to differentiate dramatically each of these poems from the other three, and it is in this that the utility of the procedure resides.

The elementary statistical measure to be employed here is the coefficient of correlation (also explained in any text book on statistics). As will be seen, the coefficient of correlation is more informative than any of the figures thus far submitted, but it requires accordingly greater labor to derive it, for it involves an inspection of each individual line of verse (a task that can be programmed without excessive difficulty). Further, some care must be used in its interpretation since, as one textbook puts it : "... the correlation coefficient is the best known and probably the worst abused measure of association known to statistics." (1)

---

(1) I.M. Craddock, *Statistics in the Computer Age* (English Universities Press, London, 1968) 40.

Since I have no intention of abusing this statistical measure, the following explanation may be somewhat tedious. To begin, a major abuse to be avoided is confusion between description and causal explanation. The coefficient of correlation measures association, but this word "association" is not to be taken in a causal sense; the word is used to name a numerical or quantitative relation existing between two measurable or countable variables. Let us call these variables X and Y. If X increases as Y increases, than there is a positive correlation between X and Y. If X decreases as Y increases, then the correlation is negative. If an increase or decrease in X is not accompanied by a corresponding change in Y, then there is no correlation between them. For example, if X is the side of a square and Y is the diagonal of the same square, then any increase in X is accompanied by an increase in Y and there exists a positive correlation between these two variables. Further, one may express this relation in the form of a linear equation ($y = \sqrt{2X}$), that is, if all the points satisfying this relation are plotted on a grid with Cartesian coordinates, they will fall upon a single straight line, which slopes upward as one moves from left to right. (See graph 1)



The above is the description of a relation whose coefficient of correlation is +1 (if one variable decreases as the other increases, e.g., $y = -\sqrt{2X}$, then the line slopes downward as one moves from left to right and the coefficient of correlation is -1). Please note that a description of the above situation in causal terms is not really appropriate. In a sense, a change in X "causes" a change in Y, but just so a change in Y "causes" a change in X.

This interchangeability of cause and effect is generally true for any relation which is expressed in purely mathematical terms. Nevertheless, there are situations in which it makes sense to speak of cause and effect, and these relationships may also be expressed mathematically, but only in a somewhat analogical or conventional fashion. We must therefore turn to those cases where the coefficient of correlation is not +1 or -1. By definition or convention, when the points satisfying a relation between X and Y are plotted upon a Cartesian grid and they do not fall upon a single straight line, then the coefficient of correlation has a value which lies somewhere between +1 and -1 (except for a special case which will be discussed below).

Let us now imagine a physical case where it is reasonable to speak of cause and effect, e.g., the diet of a cow and her milk yield, where it seems reasonable to say that the diet (x) has effects upon the milk yield (y) but not vice versa. In such a case we shall call x the independent variable, and y the dependent variable.

Now suppose that 1 unit of x results in 1 unit of y, 2 units of x produce

3 units of y; and 3 units of x produce 4 units of y. These paired values may be expressed in tabular form :

x :	1	2	3
y :	1	3	4

The simplest mathematical expression which will incorporate these three paired values is $2y = 7x - x^2 - 4$, and this is not a linear expression. If plotted on a Cartesian grid, these points will not fall on a straight line. However, for the given values of x, it is possible to find that linear expression which affords the best possible approximation to the corresponding values of y. Note that we make two demands upon this expression : (1) it must be linear, and (2) it must afford the best possible approximation to all the values of y for all the given values of x. Further, the meaning of "best possible approximation" must be clarified. In accordance with statistical convention, that approximation is best which minimizes the sum of the squares of the deviations between the true values of y and those indicated by the linear expression. Let us illustrate. In the present case, the requisite linear expression is $6y = 9x - 2$. Substitution of the given values of x yields the following values and deviations for y :

x :	1	2	3	
y :	7/6	16/6	25/6	(approximation)
y :	1	3	4	(actual)
	1/6	2/6	1/6	(deviation)

concept. On the other hand, the foregoing should not obscure the very real utility of regression analysis as an instrument of statistical description.

In order to carry this discussion further, we must now consider the sort of case in which more than one cow is involved. Instead of cows, however, we discuss, as an example, the relation between the number of words (x) and the number of syllables (y) in a line in *Eclogue 9,* always with the reminder that the previous remarks on causality continue to apply. The mathematical analysis will not tell us that the number of syllables is "caused" by the number of words, or vice versa. Further, we shall illustrate the calculations whereby the regression lines are derived, but since, despite appearances, this is not a textbook on statistics, the mathematical justification for these calculations will not be fully explained

To begin, a two-dimensional table is constructed in which is recorded the number of lines in the poem which possess a particular combination of words and syllables. These numbers are recorded in Table 10 below. Note that the vertical columns are labelled with the number of words (x) and the horizontal rows are labelled with the number of syllables (y).

TABLE 10

Eclogue 9

		Number of words (x)					
		5	6	7	8	9	10
Number of syllables (y)	17	0	0	2	2	0	0
	16	1	3	7	4	2	1
	15	3	7	12	7	2	1
	14	1	5	4	1	0	0
	13	0	1	0	1	0	0

Thus, there are 12 lines which contain both 15 syllables and 7 words, 2 lines which contain both 16 syllables and 9 words, and so on.

The regression of y on x has the form : $y = a + bx$

The formula for b is : $\frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}$

Σxy is the sum of the products of each cell's coordinates and the number it contains. Thus the contribution of the cell containing 12 to this sum is (12)(7)(15) or 1260. Σxy in this case is 7173.

$n\bar{x}\bar{y}$ is the product of the total number of lines and the mean values of x and y. In this case, $n\bar{x}\bar{y} = (67)(7.04)(15.16) = 7150.67$.

Σx^2 is the sum of the products of each x squared and the number of times it appears. Thus, the value ($x = 5$) is found 5 times (the total of the numbers in the cells in the column under that value) and the contribution of this column to Σx^2 is $(5)^2(5) = 125$. In this case, Σx^2 is 3410.

$n\bar{x}^2$ is simply 67 times the mean value of x (7.04) squared, or 3320.63. These values are inserted into the formula :

$$b = \frac{7173 - 7150.67}{3410 - 3320.63} = .249$$

Substitution of the mean values of x and y now allows the determination of a and the emergence of the regression of y on x : $y = 13.41 + .249x$.

In a similar fashion, we derive the regression of x on y, $x = .374y + 1.37$. It will be instructive to repeat Table 10 with the lines of regression superimposed.

Graph 2

Eclogue 9

Number of words (x)

Number of syllables (y)	5	6	7	8	9	10
17	0	0	2	2	0	0
16	1	3	7	4	2	1
15	3	7	12	7	2	1
14	1	5	4	1	0	0
13	0	1	0	1	0	0

The regression of y on x, y = 13.41 + .249x is indicated on Graph 2 by – – – – –. Let us make crystal clear what that line represents, since the case at hand differs from the cow example used to introduce this topic. In the introductory cow example, the regression line of y on x was used to arrive at the best possible approximation of y for a given value of x. This is still true here, but given the number and dispersion of the observations here, we must refer, not to a specific value of y, but to the mean value of y for a given value of x. For example, given the value $x = 5$, we see that

there are 5 verses in *Eclogue 9* which consist of 5 words. One of these lines consists of 14 syllables, three of 15 syllables each, and one of 16 syllables. The mean value of y is 15, but the line of regression indicates an approximate value of $y = 14.65$. Other lines might indicate a closer figure for $x = 5$, but no other line will approximate so closely each of the mean values of y for all the values of x.

The regression of x on y, $x = .374y + 1.37$ is indicated in Graph 2 by # # # # #, and it, in turn, indicates the best approximation of the mean value of x for each of the values of y. Note that the lines intersect at the overall mean values of x and y.

For most purposes (including ours) the most significant thing to note about the lines of regression is the size of the angle, which they form, for the smaller this angle is (i.e., the closer the two lines of regression come toward coïncidence), the closer the association between the variables x and y. But these lines can approach coïncidence only if, for any given value of x, the dispersion of values for y is narrow, and vice versa, and this is what the concept of association is all about. In other words, two variables are closely related if a given value of one determines a closely restricted range of values for the other. This quality of association can be measured. This measure (called the coefficient of correlation) is the cosine of the angle formed by the lines of regression, and the statistical formula for deriving this measure is :

$$\frac{\Sigma xy - n\bar{x}\bar{y}}{n\sigma_x\sigma_y}$$

We are already familiar with the numerator of this expression. The denominator is the product of the number of cases (here 64), the standard deviation of x (here 1.13) and the standard deviation of y (here 0.87).

In this case, the coefficient of correlation is .337, indicating a certain magnitude of positive correlation. (If the correlation were negative, the formula would produce a negative number). As the angle between the lines of regression decreases, the magnitude of the coefficient of correlation increases with a theoretical limit of +1 or -1. This will be the case when the two regression lines coïncide, and this will occur only when all the paired values lie on a single straight line, or in a very special sort of instance wherein all the paired values are symmetrically placed on each side of the common line of regression.

The other limiting case occurs when the coefficient of correlation is equal to zero. In such a case, the regression lines will be horizontal and vertical and they will intersect at the overall mean values of x and y, with the meaning that no magnitude of correlation exists. In other words, the relation between the variables is such that a change in one variable will not be reflected by a change in the mean value of the other.

We may now turn to a consideration of the four poems. I have calculated the coefficient of correlation for the following relations in the poems : (1) the relation of the number of words to the number of syllables in a line; (2) the relation of the number of words to the average length (in syllables) of the words in a line; and (3) the relation of the number of syllables to the average length (in syllables) of the words in a line. (It

should be noted that while the first two variables are exactly denumerable, the third, i.e., average word length is not. I have chosen to group average word lengths in intervals of 0.2 syllables with the midpoints in the intervals as follows : 1.41, 1.61, 1.81, and so on. Choice of a different interval or midpoints would affect the results to some degree).

The results are tabulated below.

TABLE 11

	Eclogue 1	*Eclogue 4*	*Eclogue 6*	*Eclogue 9*
Number of syllables and number of words	.415	.394	-.012	.337
Number of words and average word length	-.937	-.946	-.909	-.873
Number of syllables and average word length	.494	-.092	.348	.137

Consider first the correlation between number of words and number of syllables. In no case is it exceptionally large, but it is extraordinarely small in *Eclogue 6*. What little tendency there is, and there is very little, indicates that the more words there are in a verse, the fewer syllables there will be in the verse. This seems a bit perverse; indeed it is perverse, but only a tiny bit so. *Eclogue 1*, on the other hand, displays the closest approach toward correlation between these two variables, followed by *Eclogues 4* and *9*. Generally speaking there is a tendency for the number

of words to increase as the number of syllables increases, except in *Eclogue 6*.

Turning toward the relation between the number of words and the average word-length, we find a large negative correlation in every case. This is not surprising; one would expect that as the number of words increases, the average word-length will decrease. It is worth noting that *Eclogue 9* displays the weakest correlation here.

With regard to the relation between number of syllables and average word length, we again find an example of slightly perverse behavior, but this time it occurs in *Eclogue 4*. Once again, *Eclogue 1* displays the strongest correlation.

Within the limits of the conventions of statistical description, we may now answer the question put so long ago : why are the longer lines longer and why are the shorter lines shorter ? The answer differs from poem to poem. In *Eclogue 1*, the longer lines are longer both because they contain more words and because they contain longer words. This is also true of *Eclogue 9*. But in *Eclogue 1* the size of the words is slightly more important, whereas in *Eclogue 9* the number of words is the more significant factor. In *Eclogues 4* and *6*, the situation is far more striking. In *Eclogue 4*, the longer lines are those that contain more words, whereas, more often than not, the shorter lines contain the longer words. The situation is exactly the reverse in *Eclogue 6*. Here the longer lines contain the longer words, but, again more often than not, the longer lines contain fewer words.

Our findings thus far can be summarized is somewhat more impressionistic terms, but with the reminder that we are dealing with only one slight aspect of the texture of these poems.

Eclogue 1 seems to be the most regulated of the four poems; its texture is characterized by controlled restraint. There tends to be less variation in the number of words per line, and the longer lines contain more and longer words. Finally, *Eclogue 1* seems to be a model of the ideal *Eclogue*, because not only does it have qualities representative of the whole, (average-length lines, average-length words), it also presents these features in a more highly correlated fashion.

The striking feature of *Eclogue 4* is the relative shortness of its lines. Fewer syllables in the line is, of course, directly associated with a greater proportion of spondaic feet, and we may conclude that the poem has a spondaic tone, a quality that has nuances of gravity, solemnity, and highness of purpose. We may further conclude that its lack of correlation between word-length and line-length is largely due to this sustained spondaic tone. The longer lines do tend to contain more words, however, and this implies a certain uniformity in the length of words as well. These conclusions are supported by the fact (witness the standard deviations) that this poem has the least variation in the size of its lines and in the size of its words. These features taken together suggest a sustained tone maintained at a rather even level.

Eclogue 6 presents a far different picture. The lines in this poem are about average in length, but with more variation in this aspect than in

any of the other poems. The poem also has the longest words, but again with the highest degree of variation. Finally, there is that striking lack of correlation between the number of words and the number of syllables in the line, but without the justification of a sustained spondaic (or even dactyllic) tone. The adjectives that come to my mind are playful and pixy-ish. Given the lack of a pronounced dactyllic tone, one may infer a sort of light-heavy combination, perhaps a general lightness of tone punctuated at intervals by heavy and serious touches.

Eclogue 9 is the most dactyllic of the four poems and it contains the shortest words. The implication here is a general lightness of tone (perhaps a conversational quality) absent from the other poems. The comparative weakness of its correlations suggests a certain lack of control. The poem is not perverse, but rather unemphatic.

The above four paragraphs are a concession to traditional criticism, and it would be unfortunate if the importance of statistical analysis were to stand or fall because of their qualities. It is quite obvious that statistics alone are not enough, and we shall not belabor the point.

Thus far we have applied some statistical measurements which are very well known and which are generally used to describe the over-all characteristics of a group or aggregate. We have dealt with the words and syllables in each line of the poems in exactly the same way that we might treat the heights and weights of each individual in a group of men. But unlike measuring a group of men where the order in which the individuals appear is usually irrelevant, a poem is also a sequence or series of events, and the

ordering of the lines is clearly a factor of interest. These are, however, no succinct and meaningful statistical descriptions of a sequence which are as widely known or as universally accepted as the measurements used thus far. There are relatively advanced mathematical methods for the description and analysis of a series, but, for the most part, the mathematical expressions which characterize a sequence of any complexity are more fearsome to the neophyte than contemplation of the series itself. Besides, we are seeking a statistical rather than exact description of a series, that is, we seek to describe the salient characteristics of a series rather than an expression which will enable us to reproduce the series. In exactly analogous manner, the mean and standard deviation describe salient characteristics without enabling us to produce a listing of the specifics that make up the aggregate.

We begin with a simple example. Consider the following two sequences :

(a) 7, 8, 6, 7, 8, 6, 7
(b) 7, 6, 8, 7, 6, 8, 7

If taken as aggregates, the above two sequences are alike, but as sequences, one is the reverse of the other. Sequence (a) has 4 upward steps consisting of one unit each, and 2 downward steps consisting of two units each; the reverse is the case in sequence (b). We may immediately generalize that for any sequence beginning and ending with the same value, the product of the mean upward step and the number of such upward steps is equal to the product of the mean downward step and its number of occurrences. Here, very simply, 4 x 1 = 2 x 2.

Since the sequences with which we shall be dealing can never differ very much in their initial and final values, a statement of the mean upward and downward steps will also characterize their relative frequency. Again, to restate the obvious, an upward step in one direction is a downward step in the reverse direction, and hence it makes a difference in which direction one is going. Happily this is also the case (*pace* Plato, *Phaedrus* 264D) with literary compositions, and in every case we shall be moving from the beginning toward the end. As a matter of convention, we shall prefix an upward step with a plus sign (+) and a downward step with a minus sign (-).

Turning now to a specific case, if we note the series or sequence of words per line in *Eclogue 4*, we shall find that the mean steps are +1.43 and -1.65. We may immediately conclude that there are more upward steps than downward steps, and this is, in fact, the case. The respective figures are 26 and 20. If we do the multiplication (1.43 x 26 and 1.65 x 20) the products are 37 and 33, a discrepancy which is explained by the fact that the poem begins with a line of 5 words and ends with a line of 9 words.

Let us now consider the critical significance of the mean steps +1.43 and -1.65. Once again, one must beware of grandiose claims. I believe that we have grasped here one tiny corner of a critical quality which may be termed pace or flow. To put it in statistical terms, I believe that the correlation between the changes in the number of words and that total grand esthetic profusion of ideas, meanings, conceits, moods, and nuance is low but not zero, and that the changes in the word-rate make some small contribution to the total impression that we receive. Again, we shall not belabor the point.

In more impressionistic terms, *Eclogue 4* is characterized by a slight tendency to increase the number of words in a line very gradually, and to punctuate this general tendency with rather more violent decreases. A pictorial image would be that of traversing a landscape made up of slowly rising ridges followed by rather more precipitous drops.

To go on with the analysis, we have noted that *Eclogue 4* contains 26 upward steps and 20 downward steps. This leaves a remainder of 16 points where no charge takes place, i.e., where one verse is succeeded by a second verse with the same number of words. Such points, where no change occurs, constitute a serious problem for the adequate description of a sequence. Consider a few examples :

(c) 7, 6, 6, 6, 7

This sequence contains one upward step, one downward step, and two points where no change occurs. There can be, to my mind, no question that this sequence differs from

(d) 7, 6, 7

although both contain two steps of the same magnitude and in the same order, i.e., the mean steps in each case are -1 and +1.

However, one may contend, somewhat subjectively, either that sequence (c) is more gradual and therefore less striking than sequence (d), or one may equally well contend that the steps in (c) are more striking because

they occur in greater isolation. Further, to compound the problem, it may be contended that the points of no change accentuates either the step occurring before it or the step that occurs after it (I tend myself to lean toward the latter). Given such scope for subjectivity, I have arrived at the following conventional but neutral method of stating the difference between these two sequences, a method which is not completely satisfactory, but which avoids some of the major difficulties. Upward and downward indices are constructed by multiplying the mean upward and downward steps by the percentage of times that change of any sort occurs. Thus, for example, the upward and downward indices for sequence (d) are +100 and -100, but for sequence (c) they are +50 and -50. Unhappily, these indices will not distinguish between the following two sequences :

(e) 6, 7, 8, 8, 8, 8, 8, 7, 6
(f) 6, 7, 7, 8, 8, 7, 7, 6, 6

We are left with indices of +50 and -50 in each case. Furthermore, the index conflates mean change with frequency of all change, with the result, for example, that the following two sequences will also not be distinguished :

(g) 5, 6, 7, 8, 9
(h) 5, 5, 7, 7, 9

Here both sequences have an index of +100 (+1x100, +2x50). These are limitations which must be kept in mind when reference is made to the

indices. On the other hand, the indices do preserve the distinction between upward and downward change, they do take account of the comparative presence or absence of change, and they do constitute a measure which may be consistently applied. As so often in statistical description, convenience must be balanced against loss of information. A more complete description would include the mean frequency and standard deviation of change, but I do not know how to combine all these into a single convenient index which would preserve the identity of all these measurements. The index, as it stands, is a relatively mild conflation which will allow us to make comparisons that are valid so long as it is quite clear what the index represents. With this *caveat*, we now tabulate the mean steps and indices for word changes and syllable changes in the four poems.

TABLE 12

Words

	Mean Step Down	Mean Step Up	Index Down	Index Up	Index total
Eclogue 1	-1.69	+1.45	-128	+110	-18
Eclogue 2	-1.65	+1.43	-122	+106	-16
Eclogue 6	-1.56	+1.79	-110	+126	+16
Eclogue 9	-1.88	+1.55	-149	+120	-29
Syllables					
Eclogue 1	-1.23	+1.20	- 92	+ 89	- 3
Eclogue 4	-1.11	+1.11	- 66	+ 66	0
Eclogue 6	-1.50	+1.30	-113	+ 98	-15
Eclogue 9	-1.72	+1.44	-112	+ 94	-18

The figures are not particularly illuminating as they stand but it should be noted that generally there is a tendency for downward steps to be larger than upward ones. For word changes, *Eclogue 6* is a dramatic exception, and once again one may see the anomalous behavior of this poem (as contrasted with the other three). With regard to syllables, one notes again that a change in syllable number will also change the relative number of dactyls and spondees in the verse. Hence it is worth noting that *Eclogues 1* and *4* minimize such change in almost symmetrical fashion. To put it another way, so far as syllables are concerned, *Eclogue 1* and *4* traverse a terrain whose upward and downward slopes are about the same, and hence they have a metrical texture of greater regularity and symmetry than that of the other two poems.

Sequences can exercise a horrible fascination, and there is a terrible tendency to play with them (as is possible) *ad infinitum* or perhaps *ad nauseam.* Thus one can continue by determining the standard deviation of the upward and downward steps (I shall not burden the reader, with these) and the coefficient of correlation between words and syllables so far as upward and downward steps are concerned. (Here the correlations are, as might have been expected, somewhat weaker than those in the top row of Table 11, but with the same ranking). Rather than pursue this line of inquiry, I turn at this point to a formalized consideration of how the sequences change in the course of the poems. Once again in a Peripatetic vein, we shall see what happens to these sequences as the poems proceed from the beginning through the middle to the end. With a statistical and objective brusqueness that will surely obscure real differences, I define beginning, middle, and end as the first, second and third thirds of the

poem (The number of transitions, or points of change, is one less than the number of verses in the poem. If the division into thirds is not even, the breakage is assigned to the first and second third in turn).

In the following description I shall utilize only the upward and downward indices as calculated for each third of each poem in word changes and in syllable changes. Given the pairs of observations (upward and downward index for each third) for each poem, one may chart a very simple profile for the word changes and syllable changes in each poem. As we shall see, for word changes each poem has a profile which differentiates it quite strikingly from the other three. But when we turn to the profiles of syllable-changes, we shall find that the poems (aside from a scaling factor) are very surprisingly similar. It may well be, of course, that extension of the investigation to all the *Eclogues* would result in findings that are not so elegantly distinguished, but no generalizations beyond the four poems in question are attempted here, and so far as these four poems are concerned, the description is valid.

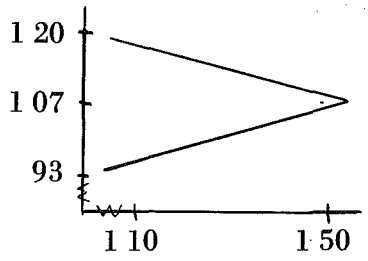
To paraphrase the Chinese, one graph is worth a thousand numbers, but it is also no pleasure to look at a printed graph which attempts to represent four poems simultaneously. As a compromise, then, I list the necessary coordinates and describe how the graph is drawn. The point determined by the coordinates for each third of a poem is located on a grid whose horizontal dimension is the downward index and whose vertical dimension is the upward index. For example, the coordinates for *Eclogue 4* are (for word changes) :

	Down Index	Up Index
Beginning	-1 09	+0 93
Middle	-1 52	+1 07
End	-1 10	+1 20

If the points representing beginning, middle and end are joined in a single line (with the point representing the middle in the middle), the result is an angle which, like an arrowhead, points firmly toward the east. See Graph 3.

Graph 3

Eclogue 4 (thirds) — Word changes



As a verbal description, the upward index for *Eclogue 4* increases slightly but consistently as we move through the poem, but the downward index increases and then decreases. The resultant profile is the arrowhead figure above, and with the usual mapping convention we say that it points toward the east (no compass mysticism is intended).

It turns out rather happily that mapping the word change indices for *Eclogues 1, 4, 6, 9* results in approximate pointings toward south, east, north, and west respectively (although the most striking aspect of the profile for *Eclogue 1* is the disparity between the first third and the very similar last two thirds). Mapping the syllable change indices however, produces a very different result. In all four cases, the pointing is clearly and decisively toward the north-east. (The tendency is most marked in *Eclogue 9*, least so in *Eclogue 6*). Again, no mystery is intended. It is merely noteworthy that in every case (there are only four !), the middle third of the poem has larger upward and downward indices than either the beginning or the end. This tendency toward greater metrical (or syllabic) change in the middle third of the poem may not be characteristic of the *Eclogues* as a whole, but it is characteristic of these four poems. This similarity of metrical practice is not accompanied by a matching similarity of practice, with regard to word changes. The coordinates are given in Table 13.

TABLE 13

Indices of change

	Words down	Words up	Syllables down	Syllables up
Eclogue 1				
Beg.	- 1 93	+ 1 54	- 75	+ 68
Mid.	- 1 02	+ 86	- 1 16	+ 1 06
End	- 92	+ 86	- 81	+ 90
Eclogue 4				
Beg.	- 1 09	+ 93	- 57	+ 69
Mid.	- 1 52	+ 1 07	- 86	+ 86
End	- 1 10	+ 1 20	- 56	+ 45
Eclogue 6				
Beg.	- 1 10	+ 1 17	- 1 09	+ 99
Mid.	- 1 17	+ 1 44	- 1 20	+ 1 09
End	- 1 02	+ 1 20	- 1 10	+ 85
Eclogue 9				
Beg.	- 1 88	+ 1 16	- 76	+ 70
Mid.	- 1 11	+ 1 07	- 1 45	+ 1 36
End	- 1 44	+ 1 38	- 1 18	+ 79

With the reminder that these indices are a conflation of mean change and frequency, we may leave this section of our analysis with the sequences of word change neatly boxing the compass and all the syllable change sequences pointing firmly in a single direction. The finding *a posteriori* is not surprising. Change in number of words is surely more intimately connected with variation in subject matter than change in total number of syllables. It is to be expected that uniformities, if detected, will tend to be found in metrical practices rather than in the number of words per line.

By now, it will have been noted that the statistical techniques thus far applied have been resolutely descriptive. There has been no mention of normal distributions, sampling, chi-square, probability, and all the other paraphernalia of inferential statistics. This omission has been intentional since there are very serious problems involved in the appropriate application of inferential techniques to the study of language and literature. Nevertheless, I shall conclude this essay in the statistical description of a literary feature by using an inferential technique in a manner whose propriety I believe I can defend.

The argument is casuistic but I believe, legitimate. The following inferential technique (known as Sherman's statistic)(1) results in a figure which may then be associated with a given level of probability or likelihood of occurrence by chance. My point may be made very summarily : the probability level constitutes a legitimate descriptive device

(1) My reference is Craddock (op. cit) 170 ff. The authoritative reference given there is *Annals of Mathematical Statistics*, vol. 21 (1950), pp. 339-361.

even if one has doubts concerning the applicability of inferential corollaries and the admissibility of the assumptions underlying such inference.

For example, one may observe an event which has a given likelihood of occurring by chance. Such likelihoods or probabilities can be measured quite precisely, and the fact that these measurements are on a probability scale means that they are standardized and comparable. The latter two qualities are quite sufficient for the description and comparison of the formal features with which we are dealing. It is most important, however, to guard the distinction between description and inference.

Thus, we shall find that 6-word lines occur in *Eclogue 1* with a regularity which would be surpassed by chance less than once in a hundred times. This is, I believe, a noteworthy fact concerning that poem. The same feature does not occur in so striking a fashion in the other three poems. Do we therefore have a right to leap to the inferential conclusion that *Eclogue 1* was not written by the author of the other three poems ? It seems obvious, at least to me, that no such inference is justified yet the application of statistics to literature seems to me to have been littered with inferential misjudgments of similar blatancy. To descend to a far more modest level, we are not even justified in inferring that Vergil paid attention (whatever that means) to the distribution of 6-word lines in *Eclogue 1*. All that we can infer is that the distribution is not due to chance (we shall see what that means).

By now it should be clear that Sherman's statistic has to do with the

intervals between the recurrences of a particular feature within a sequence. As usual, the best illustration consists of an urn filled with little balls. Our imaginary urn contains 100 balls of which 10 are red. The balls are drawn out one at a time and the intervals between the drawings of red balls are noted. We shall be very surprised if the sequence of drawings contains a succession of 10 red balls without interruption (we shall say that a tendency toward this sort of sequence is "bunched").

We shall be equally surprised if every tenth ball drawn is red (we shall designate a tendency toward this sort of sequence as "spread"). Now, one may imagine every possible different sequence of drawings that could be made from our urn with its particular mixture of red and non-red balls. We must also posit that any sequence is just as likely to occur as any other (which is how we define a random or chance event). We may then note the degree to which each sequence is "bunched" or "spread". This is done in the following manner : since our sequence contains 10 red balls, there are 11 possible intervals, and the average interval will be 100/11 or 9.1. The differences between the actual intervals and the average interval are summed regardless of sign and the whole is divided by twice the total number of balls in the urn. The resultant figure is called Sherman's statistic. Consultation of the appropriate table in the appropriate places tells us that in our particular case less than one percent of the possible sequences will have a figure lower than 0.2 (spread) and less than one-tenth of one percent will have a figure larger than 0.6 (bunched). These limits will vary according to the number of red balls involved, but generally speaking, if Sherman's statistic is greater than 0.36, the sequence has a tendency to be bunched, and similarly, if the figure is lower than 0.35, the sequence has a tendency to be spread.

If we return to the sequence of 6-word lines in *Eclogue 1*, we find Sherman's statistic to be 0.244. Given that there are 26 such lines in *Eclogue 1*, consultation of the table tells us that less than one percent of all the possible sequences containing 26 red balls (or 6-word lines) would have a figure so low, and accordingly we may conclude that the sequence is very spread indeed. If, however, we inspect the sequence of 8-word lines (of which there are 14), we find the statistic to be 0.298. This sequence is also spread, but ten percent of the possible sequences would have a lower figure. Hence we say that its probability limits are 10-50 and we conclude that the tendency toward being spread is not so marked as in the case of 6-word lines. This is about as far as I think we should go at this point, but some of the inferential implications need discussion. We may, if we wish, go on to say the following : if we assume (or hypothesize) that the recurrence of lines containing a particular number of words is random (i.e., that any sequence is a likely as any other to occur), then the case of 6-word lines would cause us to reject that initial assumption at the 2% level of significance, while the case of 8-word lines would not.

But, of course, any such initial assumption is arrant nonsense. No one, I think, would ever honestly assume, for example, that there is any chance whatever that the first and only the first 26 lines of *Eclogue 1* might consist of 6 words each, let alone the notion that such a sequence is as likely as any other, but this is what is required in order to accept fully the usual brand of inferential statistics.

While it is not the case that we must reject the possibility of ever ascertaining the proper probability levels for various linguistic usages,

ascertaining such levels would be a large empirical problem with certain built-in limitations. For we must delimit a particular universe (say, all the hexameter verses ever produced in Latin) and then inspect the frequency of occurrence of any particular usage within that universe. But the delimitation of the universe cannot escape certain difficulties absent from the model of balls in an urn. Should such a universe have chronological or genre-determined limits ?

Again, we would prefer a universe consisting of all possible hexameter verses rather than one consisting simply of those that have actually been produced. We may well believe that the actual corpus of produced verses will reproduce faithfully what the universe of possible verses would be like, but it is not clear, given the inherent limitations noted above, that such a compilation would be worth the effort necessary. What we need is a linguistic model analogous in its elegance and simplicity to that of balls in an urn, but we do not yet have that.

However, this does not mean that the techniques of inferential statistics are useless for literary studies. I believe we should find some *modus vivendi* between tossing out the baby along with his bath water and stubbornly holding on to both, when they are not both tenable. The statistical bases of language and literature are not completely analogous to those in the case of balls in an urn, but, and this is the paradoxical point, there may be some purpose in a kind of pretending that they are. For I think that the following mode of description is both valid and useful : the sequence of 6-word lines in *Eclogue 1* has a degree of spread greater than that of 99% of all possible sequences. Properly understood,

we do not even need the words "random" or "chance", for we are not here relying on any of the murky implications of those terms. But we do need the mathematical result of calculating the probability level because the Sherman's statistic alone will not allow us to compare the 10 red ball sequence with the 50 red ball sequence.

Table 14 lists the probability level for the Sherman statistic for the recurrences of lines of 5-9 words and for lines of 13-16 syllables (lines of 10 words or of 17 syllables are so few that the statistic is omitted).

TABLE 14

	Eclogue 1	Eclogue 4	Eclogue 6	Eclogue 9
Words				
5	50-90	50-90	0- 1	50-90
6	0- 1	10-50	10-50	10-50
7	10-50	10-50	1- 2	2- 5
8	10-50	10-50	10-50	10-50
9	50-90	50-90	95-98	1- 2
Syllables				
13	10-50	10-50	10-50	5-10
14	10-50	2- 5	10-50	50-90
15	0- 1	0- 1	0- 1	2-10
16	2- 5	2- 5	10-50	10-50

Some conclusions which may be drawn from the above table are the following : first, it should be noted that any finding with a 50 in it is rather weak; it represents a sequence which is neither very bunched or very spread. Second, generally speaking, the tendency is for sequences to be spread rather than bunched, but there are exceptions. Third, while there are some striking contrasts which allow us to distinguish between these poems, there is at least one striking uniformity. This is the set of extreme values found for 15-syllable lines. These are the lines whose first four feet consist of two dactyls and two spondees (in any order). This is the most common sort of line as well, and seems to represent a basic sort of rhythm which reappears with great regularity in all four poems. Finally, the treatment of 5-word lines in *Eclogue 6* is simply extraordinary. In summary, one gets the subjective impression that *Eclogue 1* and *4* are highly controlled, that *Eclogue 6* is deliberately different, and that *Eclogue 9* is simply unemphatic.

Sermons of various sorts have occurred throughout this essay and a lengthy conclusion is not necessary. First, the inane remark was made above that these statistics allow us to distinguish between these poems. It is quite clear that this has never been a serious problem. The poems are similar in some ways and very different in others, and no one has ever felt otherwise. It has, however, seemed useful to me to demonstrate that these similarities and differences extend into so mindless an area as the relationships between words and syllables. But while the area is mindless, it is also quantifiable and denumerable in a way that most literary aspects (the more important ones) are not. I do not know whether one should be happy or sad about the latter point. For my own part, I tend to be

pleased (perhaps perversely) that literary and artistic achievements have refused thus far to disclose their innermost secrets to the determined assaults of scientific objectivity, and I am reasonably confident that they never will. But, in the Terentian phrase, "*humani nihil a me alienum puto*"; I do not believe that objectivity *per se* is wicked, the ingenious design of objective analysis merits its own admiration, and the greatest compliment we can pay the citadel of artistry is not to smother it with tender, loving fatuities. If we have a way of saying how something differs, and to what degree it differs, let us use it where we can, and if we can extend the area for this sort of inquiry, so much the better.

Second, an acquaintance with statistics is not yet commonly found within the repertoire of students of classical literature and it is to these that the present essay is addressed. I assume that most of what has been said will seem otiose to the statistician. I hope it will seem correct. Perhaps the reader I am addressing does not exist, but I, at least, would have appreciated an account of regression analysis addressed to me and my interests and capacities, had I been able to find one. With the advent of the computer, statistical studies of literature will undoubtedly grow in number, and it is at least as important to know what statistics can and cannot do as it is to know what the computer can and cannot do. Both must be bereft of their mysteries before they can be appropriately used.

The present essay makes no pretense at statistical sophistication; only a few of the most elementary techniques have been applied, but these applications have been used, I hope, with propriety and confidence.

It will not have escaped the reader that nothing has been said about the content of the poems. The entire analysis herein contained could have been carried out by a person who understood not a single word of Latin. This statement is not made with pride. In mitigation, I can only hope that the findings are not without interest to those who are acquainted with the content of the poems. I have no doubt that the slight formal features described here contribute their mite to the overall aesthetic effect exercised by these poems, but I do not yet know (and probably never will) how to express that grander aesthetic profusion in confident statistical terms.(1)

Nathan A. GREENBERG
Oberlin College

---

(1) The present essay has been written in Liège while I have enjoyed the gracious hospitality of Professor Delatte and the members of his staff. My thanks to them, to the Commission for Educational Exchange between the United States of America, Belgium, and Luxembourg, and to my home institution, Oberlin College.