1. *INTRODUCTION*

Quite recently (1), N. Greenberg drew the attention of the readers of this journal to Herdan's theory on the recurrence of initial phonemes (2). In that article, Greenberg accepts Herdan's basic line of reasoning but rejects his conclusions, pointing at an "elementary but serious error" in his treatment leading "to conclusions exactly opposed to what they ought to have been" (3).

As I think that there are some corrections to be made in both Herdan's and Greenberg's exposition, I will first give a sketch of Herdan's theory and Greenberg's rejoinder to it before I come to my own proposal and investigation.

2. *HERDAN*

In his book *The Calculus of Linguistic Observations* (see footnote 2) pp. 79-85, Gustav Herdan delineates the object of his investigation as

---

* I wish to thank sincerely my colleague and friend, Dr. Frans Loosen of the University of Leuven, psycholinguist and statistician, for the many fruitful discussions I have had with him on the subject. Although I profited greatly from the discussions, the responsibility for the views and meanings expressed in this article remains solely mine.

the repetition of the initial letters (4) of verse lines, being one of the various kinds of alliteration, a well-known prosodic device (5). "On reading the *Georgics* of Virgil", he proceeds, "one is struck by the extremely common occurrence of couplets (triplets) of verses beginning with the same letter. Before, however, jumping to conclusions about such repetitions of initial letters being a deliberate product and thus being a characteristic device of the poet's style, one must make certain that the repetitions are not due to chance" (6).

Herdan then presents a method to establish those chance repetitions. His reasoning comes down to the following (7) : Suppose we consider the occurrence of a particular unit, e.g. the initial letter A at the beginning of verse lines in Virgil's *Georgics*. Let us call the probability of occurrence of this unit p; p is, according to Herdan, the relative frequency of the particular unit in question based on all initial verse line letters. In Virgil's *Georgics* for instance, p(A) = 0.1225 as there are 2188 verse lines in the *Georgics*, 268 of which start with the letter A. The probability of non-occurrence of this phenomenon (= q) is accordingly 1-p. To know the probability for XX, sequence of the same verse line letters, the multiplication theorem of probability may be applied : XX = p.p; XIX, sequence of two occurrences of the same unit separated by one other unit, = p.q.p.; X2X, sequence of two occurrences of the same unit separated by two other units or intervals, = p.q.q.p., etc. More generally, if one knows N, the totality of the phenomena under study (for instance, in a poem of 2188 verse lines, N = 2188) and p, the probability that one particular phenomenon will occur, the expected frequency of this phenomenon in the corpus of length N would be expressed by :

$$fe_{(x)} = N.p_{(x)}.p_{(x)}.(1\text{-}p)^{r}_{(x)} \qquad (8)$$

where fe stands for expected frequency and r indicates the length of gaps between two successive occurrences of the same phenomenon. Taking, for example, the initial verse letter A in Virgil's *Georgics* and while wanting to know how many XX-series of A mere chance alone would produce, applying the above formula would yield :

$$fe_{(A)} = 2188 \times 12.25\,\% \times 12.25\,\% \times (87.75\,\%)^{0}$$
$$fe_{(A)} = 2188 \times 12.25\,\% \times 12.25\,\% \times 1$$
$$fe_{(A)} = 32.83 \qquad (9)$$

This means that given p_A and N, one might expect by mere chance alone 32.83 "A-couplets" in Virgil's *Georgics*.

By comparing graphs of observed and expected relative frequencies of some letters (A,E,IJ,Q,S,T) for several gap-lengths (ranging from XX to X8X), Herdan comes to the conclusion that "for the letters mentioned above, it is mainly the patterns XX, X1X, X2X in which the observed figures exceed by far, and significantly so, the chance frequencies. The remaining patterns are, by and large, also in excess of the chance distribution, but not very much so". (10).

Generalizing the above statement and investigation, Herdan concludes

a) that he has devised "a mathematical method for deciding whether the repetitions of initial line phonemes in a given poem are accidental or in conformity with a definite pattern" (11).
b) that the repetitions of initial verse line phonemes in Virgil's *Georgics* "cannot only be regarded as a deliberate use of the device of alliteration, but as revealing a definite *pattern* through preference

for repetition of line initials in gaps of a specified length" (12) : viz. XX, X1X and X2X.

3. *GREENBERG*

As I already stated earlier, Greenberg accepts the "basic mode of reasoning whereby expected values may be established" in his remarks on Herdan (13). That means that he accepts the formula

$$fe_{(x)} = N.p_{(x)}.p_{(x)}.(1-p)^{r}_{(x)}.$$

However, he does not agree with Herdan's conclusion "that the attention of the poet and, of course, the reader, was fixed upon the initials of successive lines and the lines distant by one or two spaces" (14).

Greenberg, quite correctly, states that Herdan has based his conclusion on a miscalculation : the observed percentages, with which he works, when he compares observed with expected relative frequencies for gap-lengths (cf. Herdan, *op. cit.,* Table 14a and Fig. 1), are based upon a consideration of gap-lengths 0 through 8 only, ignoring the possible existence of larger gap-lengths (X9X, X10X, ... X2186X).

The expected percentages (probabilities), however, are based upon the total corpus (2188 verse lines). This error leads Herdan to unjustified conclusions. According to Greenberg there are no significant differences between the (correct) observed (relative) frequencies and the expected ones. To prove so he uses a $\chi 2$-test. As he finds no values exceeding the level of

significance he has chosen (a = 0.0005) (15), he concludes that there is no reason to reject the null hypothesis and that here only chance distributions are involved. In other words, the differences one notices between observed and expected values cannot be rejected as being something other than chance variations.

4. *DISCUSSION*

Although Greenberg rightly points out an error Herdan makes when dealing with the observed relative frequencies, he passes over a more serious mistake which lies on the level of the expected relative frequencies. Indeed, the expected relative frequency (i.e. probability) for each letter is calculated from an incorrect hypothesis. Herdan takes together all the initial letters with which the verse lines begin as the corpus from which the probability of the different initial letters, which may occur, is to be established. That this choice is wrong can be easily demonstrated. Suppose there is a Latin poem consisting of ten verse lines, the first nine of which begin with an 'A'word, the remaining one with a different initial.

The observed frequencies in this case would be, (for XX, XXX and XXXX) XX 8 / XXX 7 / XXXX 6.

The corresponding expected frequencies would be

XX	$10.\frac{9}{10}.\frac{9}{10}$	=	8.10
XXX	$10.\frac{9}{10}.\frac{9}{10}.\frac{9}{10}$	=	7.29
XXXX	$10.\frac{9}{10}.\frac{9}{10}.\frac{9}{10}.\frac{9}{10}$	=	6.56

Even without applying any statistical test at all, one can notice that the differences between observed and expected frequencies are rather small, so that when applying an $\chi 2$-test one gets a $\chi 2$-value of 0.07 which has a probability of more than 90 % of being due to chance only. Consequently the null-hypothesis that a chance distribution is involved, will not be rejected here.

In Herdan's terminology this would mean that the repetitions XX, XXX and XXXX of initial line phonemes in this poem are accidental (16). However, if it is true, and one may be certain of that, that Latin verse does not obey a deterministic system which we can simplify as : $L_{i+1} = L_i$, where L stands for the initial letter of a verse and i denotes the order of the verses, it is clear that several repetitions of the same initial phoneme at verse beginnings is something very exceptional and by no means accidental. The anomaly of the example given would not be detected by either Herdan or Greenberg because they simply base their probabilities for units to occur at the beginning of a verse (here in case $p_{(A)}$) on the *actual* occurrence of those units at the beginning of verses, instead of taking into account the real probability or *potential* occurrence of those units at the beginning of verses.

In my opinion, therefore, the probabilities for letters to occur at the beginning of verses should not be based on the occurrence of the initials of *verse line words*, that is, words at the beginning of each verse, only, but on the occurrence of the initials of *all words of the text*, as every word in the text has the same chance to occur at the beginning of the verse line - especially in a language like Latin where there is a free word-order. In other words, in the Herdan-Greenberg treatment one works with probabilities of initials which actually appear at the

verse line beginnings. Consequently when one wants to know whether this actual occurrence and its patterns are accidental or not, this question cannot be answered reliably (17). It can be done, however, when dealing with the probability of the potential verse line initials, i.e. when taking the initials of the whole text and not just of the verse line beginnings into account and calculating on this basis the relative frequency for each letter (18).

A second remark concerns the method of testing the hypothesis in question. This hypothesis can be formulated as follows : "the repetitions of initial line phonemes in a given poet are accidental". In trying to reject this hypothesis, Herdan does so without the use of statistical tests, comparing only observed and expected relative frequencies in graphical form. This certainly makes it difficult to reach valid conclusions, as significance becomes a matter of personal intuition. Greenberg uses the $\chi 2$-test. In my investigation I will use another one. Indeed, although the $\chi 2$-test is an easy one to apply, it often treats phenomena so globally that they become blurred and distinctions are lost. In order to demonstrate this, let us take up the following example.

Suppose there is a poem consisting of 100 verse lines, the first ninety of which begin with the letter A, the remaining ten all having a different initial. In this case we would have the following observed frequencies :

AA	89	
AAA	88	
A1A	0	etc.

The corresponding observed frequencies would then be

$$AA : 100.\frac{90}{100}.\frac{90}{100} = 81$$

$$AAA : 100.\frac{90}{100}.\frac{90}{100}.\frac{90}{100} = 72.90$$

$$A1A : 100.\frac{90}{100}.\frac{90}{100}.\frac{10}{100} = 8.10$$

taking AA, AAA and A1A together there would be no significant values at the 0.001 level ($\chi 2 = 11.80$, significant values occur for $\chi 2 \geqslant 13.815$), so it seems that the distribution does not differ significantly from a mere chance distribution. Yet, it is very likely for A1A to show a significant difference. One would expect it to occur in 8 % of the (triplet) cases, and it does not occur at all. However A1A is drowned, as it were, by the other values in the series AA, AAA, A1A. If, on the other hand, we would take another phenomenon into account, e.g. AAAA, the distribution would become significantly different (at $\alpha = 0.001$) from a mere chance distribution. To avoid the blurring effect that one gets when one takes several phenomena together, in the investigation which follows I will use a test which treats several phenomena separately and consequently is more suitable for the type of investigation done here.

5. *THE RECURRENCE OF INITIAL GRAPHEMES (PHONEMES) IN THE FIRST BOOK OF VIRGIL'S GEORGICS*

I have applied the above indicated formula, using the probabilities of occurrence for initial phonemes based on the first book of the *Georgics*. Consequently, questions like the following can now be answered :

What is the chance of finding an A at the beginning of a verse in the first book of Virgil's *Georgics* for example, when the words the author has used to write the text, are known ? The differences between Herdan's probabilities and my own are shown in the table below :

	HERDAN (N = 514)		MARTIN (N = 3309)	
	f absolute	f relative	f absolute	f relative
A	67	13.04	361	10.91
B	2	0.39	28	0.85
C	32	6.23	282	8.52
D	21	4.09	117	3.54
E	51	9.92	238	7.19
F	21	4.09	166	5.02
G	5	0.97	41	1.24
H	20	3.89	68	2.06
IJ	52	10.12	268	8.10
L	11	2.14	115	3.48
M	18	3.50	172	5.20
N	39	7.59	162	4.90
O	9	1.75	84	2.54
P	26	5.06	236	7.13
Q	27	5.25	102	3.08
R	5	0.97	104	3.14
S	48	9.34	338	10.21
T	31	6.03	203	6.13
UV	29	5.64	221	6.68
Z	---	---	3	0.09
	514	100.01	3309	100.01

Once the correct probabilities for the initial graphemes in the text are known, one can calculate for each grapheme its expected frequency of occurrence at the beginning of a verse : N (= total amount of verses). $p_{(x)}$

The expected frequency for an initial grapheme to be repeated at the beginning of a verse is consequently : $N.p_{(x)}.p_{(x)}$.

$N.p_{(x)}.q_{(x)}.p_{(x)}$ indicates, then, the expected amount of X1X cases for a definite grapheme X at the beginning of a verse. By taking the several values XX (or X1X, or X2X, etc.) for each grapheme together, the expected frequency for the phenomenon XX, (or X1X, or X2X, etc.) in general results.

In this way a second question can be asked : does the observed frequency of the pattern XX (or X1X, or X2X, etc.) differ significantly from the expected frequency of this pattern, that is, the sum of the expected frequencies for each letter of that pattern. I have done so for the patterns XX, X1X, X2X and X3X (19) in the first book of Virgil's *Georgics* in order to check both Herdan's and Greenberg's statements : the former coming to the conclusion that XX as well as X1X and X2X are typical for Virgil, that is, they differ significantly from what one would expect when this aspect of the text would be the product of mere chance; the latter coming to exactly the opposite conclusion, considering none of the three patterns mentioned to be typically Virgilian, but all accidental.

I came to the following results :

	f expected	f exp. rel.	f observed	f obs. rel.
XX	35	35/513 = 6.82	36	36/513 = 7.01
X1X	32	32/512 = 6.25	49	49/512 = 9.57
X2X	30	30/511 = 5.87	30	30/511 = 5.87
X3X	27	27/510 = 5.29	26	26/510 = 5.09

In order to test the differences between observed and expected relative

frequencies, I used the assymmetric confidence interval estimates for proportions (20), i.e. given the observed relative frequency (observed proportion) for a certain phenomenon (e.g. XX) the confidence limits (for a = 0.01) for the population to which the observed proportion belonged were established. In other words, with $a = 0.01$ there was a 99 % chance that proportions lying between the calculated confidence intervals belonged to the same population as the one to which the observed proportion belonged. Formulated otherwise, expected proportions were compared with observed proportions, when the expected proportions lay outside the confidence limits of the observed proportions, there was only one chance out of a hundred left that this was due to mere chance. In this case the null-hypothesis was rejected and a significant difference between both observed and expected proportion accepted. The following formula, was used :

$$p = \frac{P + \frac{Zc^2}{2N} \pm Zc \sqrt{\frac{P(1-P)}{N} + \frac{Zc^2}{4N2}}}{1 + \frac{Zc^2}{N}}$$

in which p indicates the confidence limit of a proportion;

P = the observed proportion (observed relative frequency);

Zc = the Z-value corresponding with a given a-level, i.c. $a = 0.01$, $Z = 2.58$;

± = with + the formula will yield the right confidence limit, with - the left confidence limit;

N = sample size (with 514 verses N for XX = 513, N for X1X = 512 etc.).

For XX the following calculation

$$p = \frac{0.0701 + \frac{(2.58)^2}{2 \times 513} \pm 2.58 \sqrt{\frac{0.0701\,(1 - 0.0701)}{513} + \frac{(2.58)^2}{4 \times (513)^2}}}{1 + \frac{(2.58)^2}{513}}$$

ultimately yields : p = 0.1005 for the right limit and 0.0496 for the left.

The confidence intervals for the observed relative frequencies under investigation are given below :

	f obs. rel.	confidence limits
XX	7.01	4.96 - 10.05
X1X	9.57	6.73 - 13.35
X2X	(confidence limits not calculated as there was no difference between f expected and f observed)	
X3X	5.09	3.07 - 8.16

From this table it becomes clear that of the phenomena under investigation only X1X shows a significant difference (at $a = 0.01$). Consequently it will be affirmed that the XX, X2X and X3X patterns show no or none other than chance variations and are thus 'accidental'. However, as the X1X pattern shows a significant difference between observation and expectation, it will be stated that the X1X pattern is a typical Virgilian alliteration pattern (21), at least in the first book of the *Georgics*.

6. *TEXTUAL ANALYSIS*

Before coming to conclusions, I would like first to come back to the text

under investigation. In this note there is no place to give a detailed textual analysis like the one S. Govaerts has given when dealing with a similar problem (22). However, when one considers the X1X pattern in the first book of Virgil's *Georgics* (23), it is striking that the pattern occurred mostly in clusters, often forming a kind of 'crossed' strophe as some of the examples show :

v 47	*illa*	v 295	*aut*	v 281	*ter*
	agricolae		*et*		*scilicet*
	illius		*at*		*ter*
	at		*et*		*septima*
v 447	*Tithoni*	v 476	*vox*	v 510	*vicinae*
	heu		*ingens*		*arma*
	tam		*visa*		*ut*
	hoc		*infandum*		*addunt*

This makes the phenomenon X1X certainly more striking and unexpected.

7. *CONCLUSIONS*

In this note I have tried to make clear that, although Herdan was right in establishing a formula to predict the chance repetitions of phonemes in the beginning of verses, he could not arrive at correct results in predicting the amount of alliterative types in texts as he used incorrect probabilities. On the other hand, when the formula was applied with correct probabilities and when observations and expectations were compared, the conclusions Greenberg comes to must be rejected as he finds only chance variation. From this investigation it becomes clear that the X1X-pattern in the

first book of the *Georgics* is a typical Virgilian type of alliteration, in the sense that it is a pattern which significantly differs from what one would expect when taking chance alone into account. In this respect I can partly agree with Herdan who noticed that the amount of X1X-patterns in Virgil's *Georgics* was not accidental but due to Virgil's deliberate choice. However, this deliberate choice should not be extended - at least in the first book of the *Georgics* - to the XX and X2X patterns as Herdan claims.

Institute of Applied Linguistics
UNIVERSITY OF LEUVEN (Belgium)

Willy MARTIN

NOTES

(1) N. GREENBERG, *Line initials in the Georgics,* in *Revue de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur,* 1972, N°4, pp. 55-58.

(2) G. HERDAN, *The Calculus of Linguistics Observations,* 's Gravenhage, *Mouton,* 1962, pp. 79-85.

(3) N. GREENBERG, *op. cit.,* p. 55.

(4) Herdan speaks mostly about phonemes when he actually means graphemes or letters. I will make no strict distinction either, so as not to hamper the discussion needlessly.

(5) G. HERDAN, *op. cit.,* p. 79.

(6) G. HERDAN, *op. cit.,* p. 80.

(7) G. HERDAN, *op. cit.,* pp. 80-85.

(8) Actually, Herdan gives the formula in the form of $p_r = p.(1-p)^r$, to denote the probability that a certain gap will occur. If we know the probability that a certain phenomenon will occur in a certain corpus (= N.p) we can calculate the frequency of this phenomenon to be repeated with a certain gap length as $(N.p) .p.(1-p)^r$.

(9) Cf. also N. GREENBERG, *op. cit.,* p. 56.

(10) G. HERDAN, *op. cit.*, p. 85.

(11) G. HERDAN, *op. cit.*, p. 85.

(12) G. HERDAN, *op. cit.*, p. 85.

(13) N. GREENBERG, *op. cit.*, p. 55.

(14) G. HERDAN, *op. cit.*, p. 85.

(15) This is an unusually high level. It is very interesting in case we can reject the null-hypothesis, then there remains only a probability of 0.0005 that possible differences will be due to mere chance; however, it also means that when one cannot reject it one does not accept differences as being significant, as long as they do not amount to or surpass this very high significance level. Taking this into consideration, in our investigation a will be taken at the 0.01-level.

(16) G. HERDAN, *op. cit.*, p. 81.

(17) That means that the answers given will always be restricted by the fact that one assumes that one knows beforehand what selection of initial verse phonemes the author has made. Consequently the 'nature' of this selection can never be questioned.

(18) It is assumed then that the distribution of initials in the whole text is a much better approximation to a random distribution

of initials than the distribution of verse line initials. Yet an author using a particular alliteration pattern in the text very frequently could disturb this random distribution. In this case the difference between expected and observed frequencies for verse line initials should result in a significant deficit for the observations (of the particular pattern in question), which is not the case in the text under investigation.

(19) I considered larger gap-lengths as losing their perceptive impact on the reader rather rapidly. This does not mean that testing other patterns would be senseless (e.g. XXX).

(20) See M.R. SPIEGEL, *Theory and Problems of Statistics*, New York, *Schaum*, 1961, p. 162.

(21) The same is stated by HERDAN in a footnote (6) on p. 80, *op. cit.*

(22) S. GOVAERTS, *Les initiales de vers chez Lucrèce et Virgile,* in *Statistique et Analyse linguistique,* Paris, P.U.F., 1966, pp. 41-49.

(23) The verses refer to the edition of H.R. FAIRCLOUGH, *Virgil I (Eclogues, Georgics, Aeneid I-VI)*, Cambridge (Mass.), 1957.