ON SENTENCE-LENGTH AS A STATISTICAL CHARACTERISTIC OF STYLE IN PROSE: WITH APPLICATION TO TWO CASES OF DISPUTED AUTHORSHIP

By G. UDNY YULE

Section I. Introductory

One element of style which seems to be characteristic of an author, in so far as can be judged from general impressions, is the length of his sentences. *This* author develops his thought in long, complex and wandering periods: *that* finds sufficient for his purpose a sequence of sentences that are brief, clear and perspicuous. Since the length of a sentence can be readily measured, for practical purposes, by the number of words, it occurred to me that it would be of interest to subject this impression to statistical investigation.

In carrying out the investigation, I met with more difficulties than I had foreseen. There are two terms used above: (1) Sentence, (2) Word. What is a sentence? What is a word, or what for present purposes is to be regarded as a word?

Sentence. Let me cite the *New English Dictionary*:

> Sentence. *sb.* 6. A series of words in connected speech or writing, forming the grammatically complete expression of a single thought; in popular use often (= Period *sb.* 10) such a portion of a composition or utterance as extends from one full stop to another. In *Grammar*, the verbal expression of a proposition, question, command, or request, containing normally a subject and a predicate (though either of these may be omitted by ellipsis). In grammatical use, though not in popular language, a sentence may consist of a single word.... English grammarians usually recognise three classes: simple sentences, complex sentences (which contain one or more subordinate clauses), and compound sentences (which have more than one subject or predicate).

From these definitions I conclude, I hope rightly, that we may drop the term "period" and use the term "sentence" to cover *any* sentence (or as I should have been inclined to write "period"), however complex and however compound in the senses defined. It is convenient to be able to avoid a term which to a statistician would generally suggest a different meaning. Now, not being a grammarian but just one of the populace, I confess that I started with the popular notion of a "sentence" in this general sense: "such a portion of a composition as extends from one full stop to another", and thought I would have nothing to do but tot up the words from full stop to full stop. The first definition, however, reads: "the grammatically complete expression of a single thought." I feel some doubts as to the "*single* thought". (Is not "I am tired and hungry" a sentence, and does it not convey two thoughts, the thought of being tired and the thought of being hungry?) But the "grammatically complete

expression" surely is essential to make a word-series a sentence; the word-series must be what Webster calls a "sense unit", and the trouble is that, especially in older works, "a portion of a composition" which "extends from one full stop to another" is often *not* the grammatically complete expression of anything. When the author or compositor has used punctuation in this fashion it is no longer possible simply to add up words from one full stop to the next, paying little or no attention to sense: it is necessary for the reader frequently to pull up and ask himself if the words just read do or do not form a sentence, and if they do not, what are in fact the limits of the sentence within which they must be assumed to lie. I need hardly point out how much this increases labour, and even, if the sentences are very long and complicated, brings in largely the element of personal judgement. Two readers, at least unskilled readers like myself, may well differ as to where a given sentence terminates.

Here is quite a simple illustration of the difficulty from a modern essay on *The Politics of Burns* (ref. 1, at end of paper):

> There are several points here all at once calling for notice, and seldom getting it from friends of the poet:
> The extraordinary talent for history shown by Robert Burns.
> His attention to British History in preference to Scottish.
> The originality of his views.

In this passage there are four word-series, the first divided from the second only by a colon (though the second begins with a capital letter), the second divided from the third, and the third from the fourth, by full stops. But neither the second, nor the third, nor the fourth word-series is a grammatically complete expression. The whole passage must be taken together, as it seems to me, as one single sentence. I am of course simply illustrating my difficulty, not criticizing the punctuation.

On the other hand, where an author has written a very long and meandering sentence, a question may well arise between two different readers as to whether a halt should not be called in the middle, and a full stop entered where author or compositor has placed only a colon.

I say author *or* compositor, for it must not be assumed that one is necessarily laying sacrilegious hands on the deliberate construction of the author himself. "So far as punctuation is concerned," says McKerrow (ref. 2), "there seems very little evidence that many authors exercised any care about it whatever. After all, even at present, few authors trouble to punctuate their MSS. with any care or consistency. Such punctuation as is found in ordinary MSS. of the sixteenth and seventeenth centuries is indeed most erratic and seldom goes beyond full stops at the end of most of the sentences and some indication of the caesura in verse." I had, before I started the present work, expected that this comment would apply much more to intermediate punctuation than to full stops, trusting that authors would at least insert "full stops at the end of *most* of their sentences"

But that it applies to both was enforced on me by different versions of the short tract by Gerson, *De Meditatione Cordis*, in the edition of his complete works that I used (see below section III and ref. 9) and in four editions of the *Imitatio Christi* on my shelves. The versions differed, not only verbally, but also as regards full stops. If punctuation, even as regards full stops, is largely the work of the compositor, there need be no hesitation in overriding them if necessary: indeed, the use of personal judgement seems unavoidable.

Let me add that at first I by no means realized the full extent of this difficulty, and when I did often felt myself horribly incompetent to deal with it. I am sure my final decisions could often be contested, and were not infrequently inconsistent with one another. But after all difficult cases are but a small proportion of all sentences in most writers and, if only as an exploratory piece of work, I hope the investigation may still retain interest and value.

Word. Compared with the difficulties as to the sentence, the difficulties concerning words are really of a minor kind. One large class is indicated by the lines of Calverley:

> Forever; 'tis a single word!
> Our rude forefathers deemed it two:
> Can you imagine so absurd
> A view?

Our rude forefathers also wrote *it self*, *any where*, *every where* and so forth, where their rude descendants write *itself*, *anywhere*, *everywhere*. How shall we reckon such expressions? It is best, I think, to follow modern usage and I generally endeavoured to do so; but in rapid counting it is very easy to make a slip. Hyphened words present the same sort of difficulty. *Law-courts*, *china-manufacturer*, *news-journal*, *well-earned*, I would count as two words each; *out-of-the-way* as four: but *co-acervation*, *contra-distinguish*, *tri-syllabic*, *pre-disposed*, *re-produce*, as one each. A *something-nothing-every-thing* (Coleridge) presents a special problem: I think it should be three words. But how many words is *matter-of-factness*? Coleridge calls it *a* word, "an uncouth and new coined word".

Then there are abbreviations such as *viz.*, *i.e.*, *etc.* or *&c.* The first there is no reason to reckon as anything but one word. The second, third and fourth in spite of their meaning, I also reckoned as one each: eye and mind grasp them as wholes.

Finally, what are we to do with figures? Dates may occur even in literary or historical essays: any year stated in figures (1825 or 1798) I reckoned as a word. Whether days of the month ever occurred I do not recall: but I would reckon the day of the month stated in figures, as in January 10th, as a word for the month and a word for the number of the day. Any actual number if stated in figures, and such numbers are frequent of course in the work of Graunt and Petty that I have discussed, would be reckoned as one word whatever the

number. Thus 251 would be reckoned as one word and so would 3,251,452; although two hundred and fifty one would be five words, and three million, two hundred and fifty one thousand, four hundred and fifty two would be thirteen. This may seem arbitrary: but again, if the number is stated in figures eye and mind grasp it as a whole, while if in words it has to be taken word by word. For the same reason, fractions such as $\frac{3}{4}$ or $\frac{1}{2}$, which are also frequent in Graunt and Petty, were reckoned as a word each. Sums of money stated in figures, such as £1. 2*s*. 8*d*. were to the best of my recollection treated as if pounds, shillings and pence were so expressed in words—not very consistently with the principle stated above. If any matter was so full of figures that it practically ceased to be prose even in the humblest sense of that term, if for example it was set out in tabular or semi-tabular form, it was simply cut out.

In all such instances as the above I really do not think it is of very much practical consequence what rule is adopted: nor even of much practical consequence if the treatment is not always self-consistent. Sentences vary too much in length for what are after all minor errors of measurement to be of much consequence.

Quotations. I may mention in conclusion one other difficulty. What is to be done with quotations? Two cases seem clear. If the author makes a brief quotation forming grammatically part of his own sentence, he is only substituting someone else's words for his own and they must be counted in: as in Lamb's

> But I am none of those who—
> *Welcome the coming, speed the parting guest.*

If, on the other hand, the author simply quotes a complete sentence from somebody else, *that* is not the author's writing and must be omitted: as for example when the same author writes

> A *gag-eater* in our time was equivalent to a *goul,* and hold in equal detestation. — suffered under the imputation.
> —*'Twas said*
> *He ate strange flesh.*

The quotation must be dropped. But no rule can be applied strictly to living literature. Thomas à Kempis, for example, quotes the words of scripture so freely that if one cut out scriptural quotations one would eliminate a considerable proportion of his work. He has made scripture his own, and what he has written must stand as his.

A serious difficulty arises only when, say, an essayist is discussing a poet and makes a long and purely illustrative quotation. This may be of any length, and it may be so made as virtually to form part of the sentence of the critic himself, or may follow almost indifferently a colon or a full stop at the end of the critic's sentence. Quotations made in the first way, and even those made in the second way after a colon, I tended at first to include. But, on coming across *very* long

quotations, it became obvious that this was unsatisfactory, and I then adopted the easier method of simply cutting out all pages on which this source of trouble was serious. This is, I think, the best course.

SECTION II. ILLUSTRATIONS FROM BACON, COLERIDGE, LAMB AND MACAULAY

This section is in part purely illustrative, showing what sort of distributions of sentence-length we may expect, but in part is concerned with the fundamental question, how far sentence-length is really a *characteristic* of an author's style. If, that is to say, we take two lengthy passages, each containing a few hundred sentences, from a given fairly homogeneous work, will they present us with proportional numbers of sentences of each particular length in reasonably close agreement with one another? If they do not; if, although dealing with the same sort of material in the same sort of way, the author is liable capriciously to vary in the length of his sentences, sentence-length is not a *characteristic* of his style in any proper sense of the term, and one's impression to the contrary will be proved mistaken. If, however, there is reasonably close agreement, we can accept sentence-length as a characteristic. It is necessary, I think, to insert the condition that the author shall be dealing with the same sort of material in the same sort of way, since (again judging from general impressions) it seems clear that sentence-length may be affected by the author's matter as well as by his individuality: argumentative passages, for example, may well tend to longer sentences than matter purely descriptive.*

The four authors chosen as illustrations are Bacon, Coleridge, Lamb and Macaulay; and their works, Bacon's *Essays*, Coleridge's *Biographia Literaria*, Lamb's *Elia* and *Last Essays of Elia*, and Macaulay's *Essays*. The particular editions used are not probably of any importance in this instance but are cited in the references at the end of the paper. They were simply those that I happened to have on my shelves.

The fundamental tables, all in the same form and showing the numbers of sentences with 1 to 5, 6 to 10, 11 to 15 words, and so on, are given in the Appendix.

Table A gives the data derived from Bacon's *Essays*. Here, when I had got to the end of Essay XXVI, "Of Seeming Wise", I judged myself to be about half-way, and called this batch of 462 sentences sample A: I then proceeded to the end of Essay LI, "Of Faction", and as this had given me 474 sentences, or approximately the same number, I called it sample B. The total number of essays being 58, the two samples together cover almost 90% of the essays. Table A shows, in addition to the distributions for the two samples

* Compare, for example, in Hazlitt's *Lectures on the English Comic Writers*, the style of the first essay "On Wit and Humour" with that of the subsequent lectures on definite groups of writers. See also below, section IV, for some remarks on Petty.

A and B, the total distribution for the two together. From inspection it will be clear that the two samples are very concordant, though figures are inevitably slightly irregular and fluctuating. In both the frequencies increase rather abruptly in the interval 11–15; in both they reach a maximum in the interval 31–35, and then tail away very slowly indeed, so that there is a considerable number of sentences of 101–200 words in length and a few over 200. The record is a sentence of 311 words, as punctuated, i.e. from full stop to full stop. The reader will find it in the penultimate paragraph of Essay XXVII, "Of Friendship". It might well be broken up: but I do not think at this early stage I had attempted any revision of punctuation, hardly having realized the difficulty mentioned in the preceding section.

Table B gives the data from Coleridge's *Biographia Literaria*. I began at the beginning and continued to about the middle of chapter IX, when I had a batch of just over 600 (actually 601) sentences, which I judged sufficient: this is sample A. For sample B I meant to take a similar batch from near the end and began with chapter XX in vol. II, not noticing that a great part of the remainder of this volume consisted of "Satyrane's Letters". The result was that chapter XX to the end gave me only about half the number of sentences wanted, and to complete the sample I went back to the beginning of the volume (chapter XIV) and worked on from that point to about the middle of chapter XVIII. This gave me sample B of 606 sentences. Again, inspection of the table shows that the distributions for samples A and B are closely alike and somewhat different from those of Table A. The actual maximum frequency occurs earlier, at 26–30 for sample A, and 21–25 both for sample B and for the two samples together; and the distribution is less scattered, there being a smaller proportion of the very long sentences of over 100 words in length. With *Biographia Literaria* the quotation difficulty became at times acute: a page or two, or a shorter passage, was omitted here and there to evade it.

The data derived from Lamb's essays are given in Table C. Sample A was taken from *Elia* (1st edition, 1823), from the beginning to some two-thirds of the way through "Mrs Battle's Opinions on Whist". Sample B was drawn from the middle of the *Last Essays of Elia* (1st edition, 1833), starting with the essay "Detached Thoughts on Books" and continuing to the end of "Barbara S—". Once more, the general consistence of the two samples looks quite satisfactory. Short sentences are much more frequent than with Coleridge, and the greatest frequencies occur in the intervals 6–10 and 10–15, which are almost equally frequent.

Finally, in Table D we have the data from Macaulay's *Essays*. Sample A was taken from the beginning of the essay entitled "Lord Bacon" (1837): sample B from the beginning of the essay on the Earl of Chatham (1844). In this instance the two samples do not agree quite so well as in previous tables. The first three frequencies are quite concordant and agree in placing the maximum frequency

at sentences of 11–15 words. But thereafter the frequencies of sample B exceed those of sample A right up to the interval 46–50, after which the position is reversed, so that the second sample is less scattered than the first. But the difference is not great.

So far we have dealt only with the similarities and differences suggested by brief inspection of the tables, but it is desirable to summarize in terms of statistical measures. Distributions of this kind, with long tails in which rather wild outliers may occur, might, it seemed to me, be best dealt with by the method of percentiles. While therefore I have calculated the arithmetic means as the most familiar form of average, I have also given the median, and for the rest have contented myself with the lower and upper quartiles Q_1 and Q_3, the interquartile range $Q_3 - Q_1$ as a measure of dispersion, and the ninth decile D_9 as an index to the extension of the tail of the distribution. These percentiles are calculated on the usual convention that the intervals may be regarded as 0·5–5·5, 5·5–10·5, 10·5–15·5, etc., and the distribution treated as continuous.*

These constants, for Tables A–D, are given in Table I. The table brings out very well the degree of consistence of each author with himself, and his differences from the others. For samples A and B of Bacon, mean, median, lower quartile and interquartile range agree within less than a unit, upper quartiles differ by 1·5 units and ninth deciles by 2·4, no very great difference from the practical standpoint especially in the constants most affected by fluctuations of sampling. For Coleridge, the two samples differ by between 1 and 2 units in the case of mean, median and lower quartile; the upper quartiles differ by 3·3, the interquartile ranges by 2·1 and the ninth deciles by 4·2. For Lamb the differences are less than a unit in the case of mean, upper quartile and interquartile range, the difference is exactly a unit for the two lower quartiles, 1·3 units for the medians, and 3·6 units for the ninth deciles. For Macaulay the

* As offprints at least of this paper may fall into the hands of some who are not statisticians, I may be forgiven for a note of explanation. The arithmetic mean is the common form of average, the sum of the quantities to be averaged divided by their number. Given a frequency distribution, it is calculated on the assumption that all observations falling into any one interval have the mid-value of that interval, e.g. that all sentences in the interval 6–10 are eight words long: this gives quite a close approximation. The lower quartile is the sentence-length such that one quarter of all sentences are shorter and three quarters longer. But sentence-lengths are discontinuous: sentences of 25 words or less might be less than a quarter of the whole, sentences of 26 words or less more than a quarter; hence some convention is necessary if a precise value is to be stated. The convention is that given in text above, and we proceed by simple interpolation. Thus in the total distribution of Table A the total number of sentences is 936, one quarter of which is 234. The first four frequencies up to and including sentences of 25 words, or up to the conventional limit 25·5, give a total of 212, and accordingly we require 22 more. There are 85 in the next interval, which is an interval of five words, and the lower quartile is therefore approximately

$$25{\cdot}5 + \frac{22}{85} \times 5 = 26{\cdot}8.$$

The upper quartile, the value exceeded by only one-quarter of the observations, and the ninth decile, the value exceeded by only one-tenth, are similarly determined.

TABLE I

Constants for the distributions of sentence-length in samples from Bacon, Coleridge, Lamb and Macaulay (Tables A, B, C and D of Appendix). Q_1 = Lower Quartile, Q_3 = Upper Quartile, D_9 = Ninth Decile)

Constant	Bacon			Coleridge		
	A	B	Total	A	B	Total
Mean	48·4	48·5	48·5	41·2	39·5	40·3
Median	39·4	39·4	39·4	35·7	34·2	34·9
Q_1	27·2	26·4	26·8	22·9	21·8	22·3
Q_3	61·7	60·2	60·9	53·2	49·9	51·3
Q_3-Q_1	34·5	33·8	34·1	30·3	28·1	29·0
D_9	89·5	91·9	91·0	74·5	70·3	73·1
	Lamb			Macaulay		
	A	B	Total	A	B	Total
Mean	26·2	26·3	26·2	22·8	21·4	22·1
Median	18·3	19·6	19·1	18·2	18·9	18·6
Q_1	10·5	11·5	11·0	11·5	12·0	11·7
Q_3	33·3	34·0	33·7	28·2	27·5	27·8
Q_3-Q_1	22·8	22·5	22·7	16·7	15·5	16·1
D_9	57·5	53·9	54·9	44·2	39·1	40·6

constants seem almost more self-consistent than inspection of the table would lead one to expect. The differences are, for means 1·4, medians 0·7, lower quartiles 0·5, upper quartiles 0·7, interquartile ranges 1·2, ninth deciles 5·1: the lessening of the scatter has affected mainly the ninth decile. For Coleridge all the constants given are lower than the corresponding constants for Bacon, the differences being most conspicuous for the upper quartile and the ninth decile. Comparing Lamb and Macaulay, medians and lower quartiles are much the same, but Macaulay's mean, upper quartile, interquartile range and ninth decile are appreciably lower than the corresponding figures for Lamb.

We may conclude accordingly that sentence-length *is* a characteristic of an author's style. There is no discrepancy between the results of our statistical investigation and the judgement made from general impressions. Given similar material and mode of treatment, an author's frequency distribution of sentence-lengths does remain constant within fairly narrow limits. At the same time, it must be admitted, the limits cannot be precisely defined. In case of dispute as to whether two works are or are not by the same author, a judgement based on frequency distributions of sentence-lengths for the two must in the end be a

personal one, and founded on such differences as are observed between samples from works known to be by the same author. Hence the importance of the illustrations that have been given.

The test is numerical, but not exact. For there can be no question of applying the ordinary tests based on the theory of simple sampling. The "samples" we have taken are in no sense random samples: they are continuous passages, or collections of continuous passages, and if (as was my practice) the lengths of sentences are written down in order as they occur it is very clear that the resulting numerical series is not a random series but a "clumped" series. Short sentences tend to occur together. The tendency is much clearer for some authors than for others and for Macaulay is a characteristic trick of style, a point being emphasized by a series of hammer-blows from sentences of very few words: for example,

> These are the old friends who are never seen with new faces, who are the same in wealth and in poverty, in glory and in obscurity. With the dead there is no rivalry. In the dead there is no change. Plato is never sullen. Cervantes is never petulant. Demosthenes never comes unseasonably. Dante never stays too long.

Or again,

> The two sections of ambitious men who were struggling for power differed from each other on no important public question. Both belonged to the Established Church. Both professed boundless loyalty to the Queen. Both approved the war with Spain.

It is obvious that a series formed from the lengths of such sentences is not a random one and that consequently differences between samples taken as we have taken them may greatly exceed the limits of *simple sampling* without, for practical purposes, being of any real significance. The differences between the upper quartiles and between the ninth deciles of the two samples from Coleridge, for example, are 10 or 11 times the standard errors, but cannot be regarded as very material.

One point regarding the form of these distributions may be noted as of interest to the statistician. They are not of the Poisson type but of the type in which the square of the standard deviation largely exceeds the mean. The following are the figures for the total distributions, the unit being a word:

	M	σ^2	σ
Bacon	48·45	1048·22	32·38
Coleridge	40·34	677·10	26·02
Lamb	26·25	514·14	22·68
Macaulay	22·07	230·04	15·17

I now pass on to an application of the method to a case of disputed authorship

Section III. The authorship of the *De Imitatione Christi*: Thomas à Kempis and Gerson

Although the old controversy as to the authorship of the *Imitatio* still continues, and only last year a translation from Netherlandish texts was published in America (ref. 7) attributing it to Gerald Groote, the founder of the Brothers of the Common Life, few I believe will not hold it to have been definitely settled in favour of Thomas à Kempis. That certainly is my belief. Any reader who wants to know more of the evidence will find a brief summary in ref. 11, or a more detailed treatment in refs. 10, 12 and 13. If this does not suffice he can follow up De Backer's bibliography, ref. 14. But I thought it would be of some interest to see what results the present method would yield when applied to investigate the respective claims of Thomas à Kempis and one of those to whom the authorship was formerly attributed, Jean Charlier de Gerson, Chancellor of the University of Paris. That Gerson could have written the book seems plainly impossible since, apart from all questions of style, it was clearly written by one who was living the monastic life; but many early editions of the book bear his name, and in others the *Imitatio* is followed by Gerson's tractate *De Meditatione Cordis* almost as if it formed part of the same work.

Since many works of Thomas are extant, admitted as such even by those who deny his authorship of the *Imitatio*, we can deal with two problems: (1) does the distribution of sentence-length in the *Imitatio* resemble that in (other) admitted works by Thomas, or no?; (2) does the distribution of sentence-lengths in the *Imitatio* resemble that in the works of Gerson?

The edition of Thomas's works that I used was that of Pohl (ref. 8). In this edition the four books of the *Imitatio* are (to retain the usual numbering) placed, as in the Brussels autograph MS., in the order I, II, IV, III. The four books are of very different lengths, covering in this edition some 51, 29, 47 and 120 pages respectively. To get a sample fairly distributed over these books, in rough proportion to their respective lengths, I took ten subsamples of about 120 sentences each as follows: Lib. I, two, from the beginning and from near the end; Lib. II, one, from about the middle; Lib. IV, two, from the beginning and from near the end; Lib. III, five, distributed through the book. The subsamples from books I, II and IV together form sample A of Table E in the Appendix, and the five from book III, sample B. Sample B contains a rather higher proportion of very short sentences, but otherwise A and B are reasonably concordant. There was comparatively little trouble with the sentence-problem: Thomas was careful in punctuation, which may be taken as his own. But one point may be noted which occurs both in the *Imitatio*, in the miscellaneous works and in Gerson: it is a question arising from the punctuation of quotations or sayings. The following from the *Soliloquium Animae* will serve as an illustration:

Caeli dixerunt. Pertransivit nos et ascendit: invaluitque supra nos. Terra respondit. Si caeli caelorum non capiunt: nolite me interrogare. Stellae cecinerunt: tenebrae sumus et non lux si illuxerit. Mare contremuit et ait. Non est in me: et abyssus ignoravit.

Here there is a full stop after *dixerunt, respondit, ait,* before the words spoken are given, although after *cecinerunt* only a colon. In all cases, it seems to me, the words spoken or quoted should be counted in with the preceding words as if there was only a colon. Further, in Lib. III I have to confess to a piece of carelessness. A number of chapters in this book begin with the vocative "Fili." followed by a full stop. This should, I think, clearly be counted with the words following: in a translation it would be followed only by a comma. But at first I had entered the word as a one-word sentence, and did not realize that the point was important since this introduction was frequent. To have left things as they were would have created a misleading number of one-word sentences: to have revised the numbers of words in all the initial sentences of the chapters affected would have entailed more labour in altering tables than I was inclined to undertake. Finally, I simply struck out all these occurrences of initial "Fili", of which there were sixteen. Sentences in the *Imitatio* being very short, my original distributions were booked up ungrouped, and this made the number of "1's" very conspicuous.

The sample to represent the miscellaneous admitted works of Thomas à Kempis was similarly made up from ten subsamples of about 120 sentences each taken from the following:

(1) *De tribus Tabernaculis.*
(2) *Epistula ad quendam Cellerarium.*
(3, 4) *Soliloquium Animae.*
(5) *Meditatio de Incarnatione Christi.*
(6) *Sermones de Vita et Passione Domini.*
(7) *Hortulus Rosarum.*
(8) *Vallis Liliorum.*
(9, 10) *Sermones ad Novicios.*

The first five form sample A and the second five sample B of Table F. Sample A in this instance has more very short sentences, of ten words or less, than sample B, but the two are otherwise very much alike, and also resemble the distributions of Table E for the *Imitatio.* More exact comparison by the means, quartiles, etc., may be postponed till we make the summary comparison with the works of Gerson also. It is a small matter, but it may be mentioned that the "texts" of sermons were omitted.

The edition of the works of Gerson that I used (ref. 9) is in four parts folio, and a selection for a sample had to be made from this rather appalling mass, a duty which could have been better performed by someone less ignorant of his work than myself. I tried to scatter the ten subsamples of about 120 sentences well over the four parts, to avoid matter that seemed hardly continuous prose or very exceptional in style and to choose matter that, in title at least, might not be too remote from something that Thomas might have treated. To reject something as "exceptional in style" may seem a dangerous proceeding, but I have in mind actually only one particular rejection, that of *De Modo Vivendi*

Omnium Fidelium. I put this down at first from its title but threw it out after examination. It consists of a series of brief rules, stated in curt sentences, after this style:

> Regula virginum. Non sint loquaces, sed simplices corde et habitu. Ad virginitatem matris Christi cogitent et eam diligant. Choreas vitent. Inter iuuenes non sedeant, nec se ab eis palpari permittant. Non ament aliquem illicito amore. Adulatores neque adulatrices recipiant nec audiant. Orationes libenter dicant. Sordida verba et inhonesta fugiant.

I hope it will be agreed that this is not normal prose—there is no continuity of thought nor development of ideas—but an exceptional *tour de force*, and was legitimately rejected. My subsamples were taken from the following:

(1) *Sermo factus in die circumcisionis Domini coram Papa apud Tarasconem.*

(2) *Tractatus contra sectum flagellantium se.* (A bad choice, as it is impossible to imagine Thomas à Kempis choosing such a subject.) As this proved too brief to give 120 sentences, sufficient was added from *Tractatus de probatione spirituum.*

(3) *Tractatus de parvulis trahendis ad Christum.*

(4) *Sermo de vita clericorum.*

(5, 6, 7) *De consolatione theologiae.* This is modelled on Boethius, *De consolatione philosophiae.* The three subsamples were taken from the beginning, middle and end. Verse was of course omitted.

(8) *De meditatione cordis*: the whole. As this gave only 109 sentences, on my reckoning, the deficiency was made up on the next two.

(9) *Sermo de circumcisione.*

(10) *Tractatus de consolatione in mortem amicorum.*

The first five form sample A of Table G, the second five sample B. It will be seen that the two are almost remarkably consistent with one another. I should add that I found the sentence difficulty distinctly troublesome at times with this edition of Gerson: full stops seem used too frequently and other punctuation marks inadequately. This impression was confirmed by the comparison mentioned in section I.

Finally, I decided to try an experiment with a different technique, pitching on columns by a random process and taking a sample of the same number of sentences from each. The parts or volumes I was using are numbered by columns, and the numbers of columns in these several volumes are as follows:

I. 934	III. 1190
II. 878	IV. 982

a total of nearly 4000 columns. Eliminating for simplicity the last 191 columns of Part III, any column can be specified by a number under 5000, the first digit giving the number of the Part, the last three digits the column; thus 2625 gives col. 625 of Part II, 4063 col. 63 of Part IV. Sequences of four consecutive numbers beginning with a 1, 2, 3, or 4 were then extracted from Tippett's *Random Numbers* and these taken as determining columns for samples. Numbers beyond the limits given above for Parts I, II and IV were simply dropped. But

numbers might also be rejected for other reasons: (1) the column might be verse; (2) it might contain matter not by Gerson at all, or only doubtfully by him; (3) the matter might be deemed otherwise unsuitable, i.e. hardly ordinary prose (cf. the rejection on the first sampling). I found it in fact quite impossible altogether to avoid the element of personal judgement and doubt now if it was desirable to attempt it: the point is discussed at the end of section IV. Relatively little was, however, rejected under the last head and the ground covered was, I think, more varied than before. When the column was fixed, I started with the first sentence beginning therein and continued straight ahead until 20 sentences had been counted. Samples A and B of Table H are therefore founded on 30 such "random passages" each, and the total column on 60 "random passages". If the "total" columns of Tables G and H are compared, it will be seen that they are closely similar.

If now Tables E and F for the *Imitatio* and the admitted miscellaneous works of à Kempis are compared with the Tables G and H for Gerson, it will be seen that there are very considerable differences, especially in the numbers of long or moderately long sentences, e.g. of more than 50 words. In Tables E and F these number 15 and 22 respectively; in Tables G and H they total to 68 and 66. For facility of checking, frequency distributions were booked up in the subsamples of about 120 sentences, and it is natural to enquire how far such small subsamples show consistent differences: it is obvious that no high degree of consistence is to be expected. The following are the numbers of sentences of 51 words or more in the subsamples of à Kempis and Gerson respectively, ranked in order of magnitude:

à Kempis: 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 4, 5, 6, 7.
Gerson: 1, 2, 2, 3, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 9, 10, 11, 11, 12, 13.

The upper quartile for Thomas à Kempis is 2·5, and this is exceeded by 17 of the 20 subsamples for Gerson. Seven of the subsamples for Thomas have no sentences at all of such a length: there is no subsample from Gerson without at least one. In both the range of variation exceeds, as one would expect, the value that would be given by the theory of simple sampling. On that theory the variance should be approximately equal to the mean, but the means and variances are:

à Kempis: M, 1·85; σ^2, 4·33
Gerson: M, 6·70; σ^2, 11·61

Roughly, fluctuations of simple sampling account for about half the variance in each case.

The complete comparison by means, quartiles, etc. is given in Table II. Comparing first the constants for the miscellaneous works of Thomas à Kempis with those for the *Imitatio*, and looking at the columns for samples A and B in both cases, we see that the values of the means overlap, that for sample A of the

TABLE II

Constants for the distributions of sentence-length in samples from the Imitatio Christi, *from Miscellaneous admitted works of Thomas à Kempis, and from Gerson. (Tables E, F, G and H of Appendix. Q_1 = Lower Quartile, Q_3 = Upper Quartile, D_9 = Ninth Decile)*

Constant	*Imitatio Christi*			à Kempis: Misc.		
	A	B	Total	A	B	Total
Mean	17·0	15·4	16·2	16·6	19·3	17·9
Median	14·0	13·6	13·8	13·8	16·4	15·1
Q_1	10·6	9·5	10·1	9·7	11·9	10·6
Q_3	20·7	18·4	19·3	20·8	23·9	22·4
Q_3-Q_1	10·1	8·9	9·2	11·1	12·0	11·8
D_9	28·6	26·0	27·7	29·3	32·5	31·0
	Gerson: Selected			Gerson: Random		
	A	B	Total	A	B	Total
Mean	23·5	23·4	23·4	23·5	22·0	22·7
Median	19·4	19·9	19·6	19·3	18·4	18·9
Q_1	12·5	12·6	12·5	12·0	11·4	11·7
Q_3	32·0	30·4	31·3	30·9	27·9	29·5
Q_3-Q_1	19·5	17·8	18·8	18·9	16·5	17·8
D_9	45·3	43·1	44·0	43·5	43·5	43·5

Miscellanea lying between the two values for the *Imitatio*. The values for the median and for the lower quartile overlap similarly. For the upper quartiles, the lower value for the Miscellanea, viz. 20·8, only just exceeds the upper value for the *Imitatio*, viz. 20·7; and there is a similar but slightly greater difference in the case of the interquartile range and the ninth decile. In no case are the differences at all large. The two tables for Gerson show a very similar degree of consilience.

But comparison of the constants for the *Imitatio* and the Miscellanea of Thomas à Kempis with those for Gerson's works shows quite a different state of affairs. For the lower quartile alone the differences are not large nor consistent, the lower quartile for sample B of the "random passages" from Gerson lying within the range of the lower quartiles for the Miscellanea of à Kempis and the *Imitatio*. All the remaining constants in the lower part of Table II are consistently larger than those in the upper part, and the differences are the more conspicuous the more the value of the constant is affected by long sentences:

it is largest (11–19 words) for the ninth decile, and next largest (4–14 words) for the upper quartile.

These results are completely consonant with the view that Thomas à Kempis was, and Jean Charlier de Gerson was not, the author of the *Imitatio*.

SECTION IV. GRAUNT'S *OBSERVATIONS UPON THE BILLS OF MORTALITY* AND THE ECONOMIC WRITINGS OF SIR WILLIAM PETTY

The problem of the authorship of the *Observations upon the Bills of Mortality* is, in all probability, of more interest to readers of this *Journal* than that of section III. At the same time it cannot be treated so completely as the problem of that section, for we have no other and admitted works by John Graunt with which to make comparison: we can only compare the one work which is generally believed to be by him with the admitted works by Sir William Petty.

The edition that I used both for the *Observations* and for Sir William Petty's writings was the convenient edition of Hull (ref. 15). Graunt gave me a certain amount of trouble in delimiting sentences, but the trouble was far more serious with Petty. I should like to quote, but the editors might reasonably object to my quoting several sentences each two or three hundred words or so in length: I must therefore merely refer readers to the original for illustrations. The longest sentence (as I reckoned it) in the *Observations* is the first part of §4, Chapter VII (ref. 15, vol. II, pp. 370–1). Here it seemed to me that the colon after "above-mentioned" on line 11 of p. 371 should be replaced by a full stop. This still leaves the sentence one of 213 words. On the other hand it appeared to me that the next following full stop between "*Annum*" and "And" on line 15 ought to be a comma, making the resulting sentence 70 words. This seemed a fairly clear case.

Take for comparison the longest sentence (again, as I reckoned it) in the samples from Petty, quite a characteristic loosely organized sequence of paragraphs in Chapter IV of the *Political Arithmetic* (ref. 15, vol. I, pp. 295–6). I allowed this sentence to begin with the words "To which purpose", the initial words in the last paragraph at the foot of p. 295, in spite of the relative adjective; but all the nine paragraphs beginning with "The value" on p. 296 had, it seemed to me, to be reckoned as part of the sentence, for the last alone possesses a verb. The result is that the sentence, on my reckoning, only stops at the words "Eighty thousand pounds" which close the paragraph towards the foot of p. 296. This is, I think, a lenient and doubtful reckoning. The first paragraph beginning "To which purpose" might well be taken as merely a relative clause properly belonging to the preceding paragraph, the sentence really beginning with the words "Now the *Wealth* of every Nation" in that paragraph, replacing the colon preceding "Now" by a full stop. This would add another 71 words to the 257 as I reckoned it in my work. Moreover, the paragraph following my

terminal limit on p. 296 leads off with "Which computation": this then might also be reckoned as a relative clause forming part of the same sentence, right down to the concluding words "Forty Five Millions", and adding yet another 105 words. On this computation then I ought to have reckoned the sentence as one of 433 words! This may sound almost incredible, but the sentence would really be no more than an expansion of a construction like this:

> Now, the wealth of a nation consisting chiefly in its share of the foreign trade of the world, we have to consider whether the English or the French have the greater *per capita* share of that trade; to which purpose I have estimated that the total value of the exports from Great Britain and Ireland, America, Africa, the East Indies, etc. amounts to some ten million pounds, a computation sufficiently justified by the Customs returns with an allowance for smuggling etc.

There is a special *type* of difficulty that occurs repeatedly, and may be illustrated by §11, Chapter VI of the *Treatise of Taxes* (ref. 15, vol. I, p. 56). The paragraph starts "The Inconveniences of the way of Customs, are, viz.", and there then follow four numbered paragraphs with different grammatical relations to the introductory clause, like this, to abbreviate greatly:

(1) That duties are laid upon [raw materials etc.].

(2) The great number of officers requisite.

(3) The great facility of smuggling by bribery, etc.

(4) The customs and duties amount to so little that some other way of levy must be practised together with it.

No. 1 obviously forms part of the sentence with the introductory clause. Nos. 2 and 3 are not sentences as they stand, and ought to have been counted in also I think, but no. 4 is an independent sentence. Actually I find that in this case I do not seem to have obeyed my own rule that a word-sequence, to form a sentence, must be a grammatically complete expression of a thought, and nos. 2 and 3 were reckoned separately: this was, I believe done in some similar cases also. Indeed judging from the few instances where I have looked again at my classification some time after the original work was done, I seem to have been usually too merciful rather than too severe in placing the limits of the sentence. Difficulties were far more frequent and more troublesome than with any author I had tackled, and made the work both tedious and unsatisfactory, for far too much was thrown on my personal judgement. Hull says (ref. 15, pp. lxvii–lxviii):

> Unfortunately the use of rash calculations grew upon Petty, and as was to be expected, he gives widely varying estimates of the same things. It must be added that he is frequently inaccurate in his use of authorities and careless in his calculations and upon at least one occasion he is open to suspicion of sophisticating his figures.

This is sufficiently severe but I would add that, in my opinion, Petty's literary style, more especially in his argumentative writing, is loose and slovenly, indeed at times hardly grammatical. It is difficult to dissociate such slovenliness in

writing from slovenliness of thought. Only in purely descriptive matter does his style take on quite a different complexion.

> They have a great Opinion of Holy-Wells, Rocks, and Caves, which have been the reputed Cells and Receptacles of men reputed Saints. They do not much fear Death, if it be upon a Tree, unto which, or the Gallows, they will go upon their Knees toward it, from the place they can first see it. They confess nothing at their Executions, though never so guilty. In brief, there is much Superstition among them, but formerly much more than is now; for as much as by the Conversation of Protestants, they become asham'd of their ridiculous Practices, which are not *de fide*. As for the Richer and better-educated sort of them, they are such Catholicks as are in other places. (*Political Anatomy of Ireland*, Chap. XII: ref. 15, vol. I, pp. 199–200.)

That is both pithy and picturesque.

So much for the difficulties; and now let us turn to the data. Graunt's *Observations* form but a slim volume, and his sentences tend to be long: omitting all prefatory matter and the appendix, and also one or two passages with tabular matter that it seemed impossible to deal with in any other way, I obtained no more than 335 sentences in all. The distribution is shown in Table J of the Appendix. To give some notion of the consistence of the style throughout, I have also broken up the total into three approximately equal subsamples. These are so small, and the run of the figures inevitably so irregular, that no very close consilience can be expected; but the degree of consistence does not seem to be at all unsatisfactory, and is particularly close as regards the numbers of longish sentences.

For facility of comparison, I thought it would be convenient to make the samples from Petty of the same size, and so intended: but, owing to a small revision made later in the Graunt table on looking through the work again, the totals for Petty are 334 against the 335 for Graunt. Sample A was taken mainly from the *Political Arithmetic*, as the work most closely associated with his name by statisticians. But this gave me only 300 sentences, and 34 were added from the *Treatise of Taxes* to make up the desired total. Sample B was taken wholly from the *Treatise of Taxes*. The distributions are given in Table K of the Appendix, and it will be seen that they are on the whole very concordant, with the exception that A shows a larger proportion of sentences of excessive length. If comparison be made with Table J it is obvious that these samples from Petty contain a very much larger proportion of long sentences than the *Observations*. There are only 17 sentences of 101 words or more in Table J, 54 and 45 sentences of 101 words or more in samples A and B of Table K. It may be added that this difference shows itself even in small subsamples. In the subsamples A, B and C of Table J there are 7, 6 and 4 such sentences. In corresponding subsamples of 111 or 112 sentences for samples A and B of Table K there are 24, 19, 11, 11, 15 and 19.

When I had got so far, I thought it would be of interest to supplement samples A and B for Petty's writings by a sample of "random passages" taken

in the same sort of way as for Gerson in section III. Hull's edition, though in two volumes, is paged continuously and runs only to 621 pages apart from appendices, index, etc.: omitting prefatory matter, the text of the first item (the *Treatise of Taxes*) does not start till p. 18. Pp. 314–438 are occupied by Graunt, with blank pages, title pages etc. I accordingly determined "random pages" by extracting from Tippett's *Random Numbers* triplets of digits beginning with 0, 1, 2, ..., 6, but not exceeding 621, and omitting numbers between the limits 000–018 and 314–438. A considerable number of the pages so given had to be struck out as either being blank pages, or containing prefatory matter, titles, contents, etc., or something obviously unsuitable such as tabular or semi-tabular matter. Very few were struck out as otherwise unsuitable, the only condition imposed being that the text should be fairly continuous ordinary prose, even though prose containing a good many figures: the limits were left as wide as possible. On each of 33 pages accepted I counted ten sentences, starting with the first complete sentence on the page and continuing till ten had been counted. On a supplementary 34th page I counted only four such sentences, so as to make up 334 sentences in all. We are dealing here with a much smaller range of numbers than in the Gerson experiment, and repetitions may occur: in fact, of the 55 numbers of three digits which were retained as lying within my limits and of which 22 were subsequently struck out as impossible or unsuitable, two occurred twice (one being amongst the subsequent rejections) and one three times. Two or three pairs might have been expected: the one occurrence of a triplet was unlikely.

The data given by this experiment are shown in column C of Table K of the Appendix. It will be seen that the first part of this distribution differs quite appreciably from the corresponding portions of columns A and B, there being a larger number of short sentences. But the "tail" of long sentences does not differ greatly, there being 40 sentences of 101 words or more in column C against 54 in column A and 45 in column B. The main source of the divergence is mentioned below, and the value of the sample discussed.

Table III gives the brief summary comparison in terms of means, quartiles etc. Taking first the medians and lower quartiles, all the three medians for Petty are higher than the median for the total of the *Observations*, which is the comparable figure based on the same number of sentences, but the median for sample C of Petty is lower than the median for sample A (based on only 111 sentences) of Graunt. A precisely similar statement is true for the lower quartiles. All the other constants, means, upper quartiles, interquartile ranges and ninth deciles are consistently higher for Petty than for Graunt, and the differences, especially for upper quartiles and ninth deciles, quite considerable. The distributions for the two authors seem to me completely differentiated: or, to put it otherwise, the results confirm other evidence that the actual *authorship* of the *Observations* is not the same as that of the economic writings of Sir William

TABLE III

Constants for the distributions of sentence-length in Graunt's Observations *and in samples from Petty's Works. (Tables J and K of Appendix)*

Constant	Graunt				Petty		
	A	B	C	Total	A	B	C
Mean	50·1	45·5	46·9	47·5	66·1	60·2	56·3
Median	45·2	38·0	37·4	40·1	56·9	51·3	44·0
Q_1	31·2	23·8	26·3	26·8	36·1	34·7	29·0
Q_3	63·3	55·5	65·5	62·3	83·2	79·0	73·7
Q_3-Q_1	32·1	31·7	39·2	35·5	47·1	44·3	44·7
D_9	85·2	85·0	85·2	85·2	126·0	109·3	110·1

Petty. Lord Lansdowne remarked, in replying to Prof. Greenwood (ref. 18, sentence quoted in ref. 19); "For literary style, neither the Observations nor Petty's writings are conspicuous, but I have yet to learn what differences can be detected between them in this respect." Sentence-length is surely one characteristic of *literary* style, and the difference seems clear. In the wider sense of *style*, the sense in which *le style c'est l'homme même*, the *Observations* seem to me to differ wholly from Petty's writings: they suggest a man of quite a different type of mind and quite a different character. The evidence from sentence-length is interesting, but adds very little.

To return in conclusion for a moment to the method of "random passages" in relation to this method of investigation, let me deal first with the reason for the divergence of sample C for Petty's writings from the two samples A and B. The latter were taken wholly from the *Political Arithmetic* and the *Treatise of Taxes*. Examining my 33 samples of ten sentences each for sample C, I found that eight (including the triplet and the pair) which were remarkable for the proportion of short sentences all came from the *Political Anatomy of Ireland*. The distribution for these 80 sentences alone is totally different from that of sample A or sample B, the constants being as follows: mean, 34·8; median, 31·2; Q_1, 24·7; Q_3, 42·2; Q_3-Q_1, 17·5; D_9, indeterminate within the blank range 59·5–62·5, say 61. Why this difference? I have already mentioned the reason and illustrated it by a quotation from this very tract. The matter is *purely descriptive*, descriptive (in the samples concerned) of the religion, diet, clothes, language and manners of the people of Ireland, and of the Government, militia and defence of the country; and when Petty has only to describe and not to argue he can apparently write like a Christian.* The *Observations* being, I think one may say, mainly argumentative, this sample of "random passages" is not properly comparable with

* Webster and the *O.E.D.* concur in classifying this expression as "Colloq. or Slang". But after all the early Christians, judging from both gospels and epistles, *did* write in short sentences.

it: it does not deal "with the same sort of material in the same sort of way" to quote the phrase from the beginning of section II. Ludicrously enough there really is no tract of Petty's in which he *does* deal with the same sort of material in the same sort of way as Graunt, so the condition is strictly impossible of fulfilment: we did our best in taking samples from two tracts that were both argumentative, and these two samples were very fairly consistent with each other.

But this result raises the whole question of method: was I right in attempting something like random sampling at all? The notion that samples ought to be random is so firmly engrained in one's mind that it seems almost sacrilegious to object to the application of the rule in a particular case. But after all the problem surely is *not* whether a tract passing under the name of Jones does or does not resemble, in this particular characteristic, a *random* sample from the writings of Brown, but samples from Brown's writings dealing, so far as possible "with the same sort of material in the same sort of way". The method of "selected samples" is, from this standpoint, entirely justified and perfectly correct. A critic may, of course, object to the particular choice of selected samples (the particular choice in this section and the last for example): but the *method* is right, and preferable to the method of "random passages" as I used it—that is to say with as little restriction as possible in regard to matter and treatment.

But there is this to be said. In the first place, used as I used it, the method does serve in some degree as a control and perhaps a warning. It brings out very well the apparent (comparative) homogeneity of Gerson's style in respect of sentence-length, and the heterogeneity of Petty's. In combination with selected samples it better exhibits all the facts. In the second place it might be used differently, just as much care being taken in deciding whether to accept or reject a passage given by the random numbers as in the case of the "selected samples", but thereby obtaining a wider range of selection.

Further, there is a danger in random sampling to which possibly I have not paid sufficient attention, the risk of bias in sampling arising from the varying lengths of sentences and the fact that the series of sentence-lengths, in order as they occur, is not a random one. To take a simple but extreme example, suppose our book consisted of equal numbers of pages containing respectively 30 sentences of 15 words each, and 15 sentences of 30 words each. Actually then the book would contain two sentences of 15 words to one of 30 words. But if we proceeded by the method used for obtaining "random passages" from Petty, taking only a sample of 10 sentences from each page determined by Tippett's numbers, we would tend to get a sample containing equal numbers of sentences of the two lengths: the number of long sentences would be overweighted. The difficulty would be surmounted if we made the sample, not a fixed number of sentences, but a fixed length of matter, say one page: or, provided the pages in the book were arranged fairly at random, by making the sample long enough to

cover a number of pages, like my subsamples of about 120 sentences. In fact of course no real case is as simple or extreme as this, and actually it will be remembered that the "random passages" sample from Petty (sample C) gave *fewer* long sentences and *more* short sentences than samples A and B, though this is no proof that it was not in some degree biased in the direction indicated. Some possible processes of sampling might easily lead to extreme bias of this type. Suppose, for example, we decided to make a random sample of single sentences, determining the page and the number of a word on the page by random numbers, and taking the sentence in which this word happened to fall. Then, it seems to me, the chance of a sentence being "caught" for the sample would be directly proportional to its length; for a sentence of 10 words would have ten chances of being caught and a sentence of 40 words forty chances. (The difficulty is closely analogous to that of determining size of family by asking casual people as to the number of their brothers and sisters.) The risk is much lessened, in my opinion, by taking longish samples and, of course, if we are mainly concerned with comparisons and not absolute figures, is less important, for the bias is unlikely to be very different in the two authors compared by the same method. The whole question of the best method to use for random sampling is, however, worth further discussion. So far as my own experience goes, however, I am inclined to prefer the method first used, the method of selected passages of considerable length.

REFERENCES

SECTION I

(1) KER, W. P. (1925). *Collected Essays*, ed. Charles Whibley, **2**, 131. London: Macmillan.

(2) MCKERROW, R. B. (1928). *An introduction to bibliography for literary students*, p. 250. Oxford: Clarendon Press.

SECTION II

(3) BACON, F. (1888). *The moral and historical works of Lord Bacon*, introduction and notes by Joseph Devey. London: George Bell and Sons.

(4) COLERIDGE, S. T. (1817). *Biographia literaria; or biographical sketches of my literary life and opinions*. London: Rest Fenner.

(5) LAMB, C. & LAMB, M. (1905). *The works of Charles and Mary Lamb*, ed. E. V. Lucas. London: Methuen.

(6) MACAULAY, T. (1888). *Critical and historical essays*. London: Longmans, Green, Reader and Dyer.

SECTION III

(7) GROOTE, G. (1937). *The following of Christ: the spiritual diary of Gerald Groote*. Translated into English from original Netherlandish texts as edited by James van Ginneken, S.J., of the Catholic University of Nymegen, by Joseph Malaise, S.J. New York: America Press.

(8) A KEMPIS, T. H. (1904–22). *Opera Omnia*, voluminibus septem edidit Michael Josephus Pohl. Friburgi Brisigavorum: sumptibus Herder.

(9) GERSONII, IOANNIS. Doctoris et Cancellarii Parisiensis (1606). *Opera*; multo quam antehac auctiora et castigatiora; inque partes quatuor distributa. Parisiis.

(10) CRUISE, F. R. (1887). *Thomas à Kempis*. London: Kegan Paul, Trench. (Part IV deals with the authorship controversy.)

(11) CRUISE, F. R. (1898). *Who was the author of the Imitation of Christ?* London: Catholic Truth Society. (A brief epitome.)
(12) WHEATLEY, L. A. (1891). *The story of the Imitatio Christi.* London: Elliot Stock.
(13) DE MONTMORENCY, J. E. C. (1906). *Thomas à Kempis: his age and book.* London: Methuen.
(14) DE BACKER, LE R. P. AUGUSTIN (1864). *Essai bibliographique sur le livre De Imitatione Christi.* Liége: Grandmont-Donders. (Nos. 3057–3301 are items "relatifs à la contestation sur l'auteur".)

SECTION IV

(15) HULL, C. H. (1899). *The economic writings of Sir William Petty, together with the observations upon the bills of mortality more probably by Captain John Graunt.* Cambridge: University Press.
(16) LANSDOWNE, the Marquis of (1927). *The Petty Papers.* London: Constable.
(17) GREENWOOD, M. (1928). "Graunt and Petty." *J.R. Statist. Soc.* **91**, 79.
(18) LANSDOWNE, the Marquis of (1928). *The Petty-Southwell Correspondence.* London: Constable.
(19) GREENWOOD, M. (1933). "Graunt and Petty, a restatement." *J.R. Statist. Soc.* **96**, 76.
(20) WILLCOX, W. F. (1938). "The Founder of Statistics." *Rev. Inst. Int. Statist.* (5), **4**, 321.

APPENDIX OF TABLES

These tables are all in the same form, showing the numbers of sentences having the length (in words) stated in the left-hand column, in a sample or samples from the source stated in the heading and more fully in the preceding text. Thus, in a sample taken from the first portion of Bacon's *Essays*, column A shows that there was only one sentence (out of 462) of a length between 1 and 5 words, 8 with a length between 6 and 10 words, 24 with a length between 11 and 15 words, and so on. Blank lines have been omitted in the tails of the tables to save space.

TABLE A

Bacon's Essays (1597–1625)

A, first half to end of XXVI. B, second half to end of LI

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	1	2	3	121–125	3	4	7
6– 10	8	8	16	126–130	2	3	5
11– 15	24	25	49	131–135	2	1	3
16– 20	22	23	45	136–140	1	2	3
21– 25	46	53	99	141–145	3	2	5
26– 30	43	42	85	146–150	—	1	1
31– 35	57	55	112	151–155	1	2	3
36– 40	38	37	75	—	—	—	—
41– 45	24	38	62	166–170	—	1	1
46– 50	31	25	56	—	—	—	—
51– 55	23	28	51	186–190	1	—	1
56– 60	25	21	46	191–195	—	—	—
61– 65	19	17	36	196–200	1	—	1
66– 70	12	13	25	—	—	—	—
71– 75	19	8	27	211–215	1	—	1
76– 80	7	11	18	—	—	—	—
81– 85	12	11	23	226–230	—	1	1
86– 90	6	7	13	231–235	—	1	1
91– 95	6	9	15	—	—	—	—
96–100	2	11	13	311–315	—	1	1
101–105	7	3	10				
106–110	9	3	12				
111–115	4	1	5				
116–120	2	4	6	Total	462	474	936

TABLE B

Coleridge, Biographia Literaria (1817)

A, vol. I to p. 134. B, vol. II, pp. 1–66 and 104–end (p. 182)

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	9	2	11	101–105	4	6	10
6– 10	21	37	58	106–110	2	2	4
11– 15	46	44	90	111–115	1	1	2
16– 20	46	49	95	116–120	5	1	6
21– 25	58	73	131	121–125	2	3	5
26– 30	64	56	120	126–130	1	1	2
31– 35	55	57	112	131–135	1	1	2
36– 40	51	52	103	136–140	—	—	—
41– 45	49	52	101	141–145	—	2	2
46– 50	39	37	76	146–150	1	2	3
51– 55	24	29	53	151–155	—	1	1
56– 60	22	23	45	156–160	—	1	1
61– 65	21	18	39	161–165	1	—	1
66– 70	20	17	37	166–170	—	—	—
71– 75	20	9	29	171–175	—	1	1
76– 80	10	6	16	—	—	—	—
81– 85	6	9	15	196–200	1	—	1
86– 90	7	7	14				
91– 95	9	4	13				
96–100	5	3	8	Total	601	606	1207

TABLE C

Charles Lamb, Elia (1823) *and* Last Essays of Elia (1833)

A, *Elia*: from beginning to middle of Mrs Battle's Opinions on Whist. B, *Last Essays*: Detached Thoughts on Books to Barbara S— inclusive

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	29	30	59	81– 85	7	6	13
6–10	115	100	215	86– 90	3	—	3
11–15	111	100	211	91– 95	5	2	7
16–20	61	85	146	96–100	2	1	3
21–25	62	56	118	101–105	3	1	4
26–30	36	46	82	106–110	1	—	1
31–35	36	46	82	111–115	1	1	2
36–40	21	29	50	116–120	1	—	1
41–45	16	19	35	121–125	1	1	2
46–50	19	16	35	126–130	1	2	3
51–55	13	18	31	131–135	1	—	1
56–60	5	6	11	136–140	2	1	3
61–65	15	11	26	—	—	—	—
66–70	2	5	7	171–175	—	1	1
71–75	7	8	15				
76–80	3	8	11				
				Total	579	599	1178

TABLE D

Macaulay

A, from first portion of essay on Lord Bacon (1837). B, from first portion of essay on The Earl of Chatham (1844)

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	26	20	46	71– 75	4	—	4
6–10	100	104	204	76– 80	4	4	8
11–15	126	126	252	81– 85	2	—	2
16–20	89	111	200	86– 90	2	—	2
21–25	82	104	186	91– 95	—	1	1
26–30	51	57	108	96–100	1	1	2
31–35	26	35	61	101–105	1	—	1
36–40	29	39	68	106–110	—	—	—
41–45	16	22	38	111–115	1	—	1
46–50	10	14	24	116–120	—	—	—
51–55	12	8	20	121–125	1	—	1
56–60	9	3	12				
61–65	7	1	8				
66–70	2	—	2	Total	601	650	1251

TABLE E

Imitatio Christi

A, from Lib. I, II and IV. B, from Lib. III

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	8	31	39	51– 55	6	1	7
6–10	142	160	302	56– 60	1	1	2
11–15	201	175	376	61– 65	1	1	2
16–20	108	129	237	66– 70	1	1	2
21–25	72	47	119	71– 75	—	—	—
26–30	33	19	52	76– 80	—	1	1
31–35	23	19	42	—	—	—	—
36–40	11	9	20	106–110	1	—	1
41–45	3	5	8				
46–50	6	5	11				
				Total	617	604	1221

TABLE F

Miscellaneous admitted works of Thomas à Kempis

For details as to the sources of samples A and B see text

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	33	14	47	51–55	3	5	8
6–10	153	98	251	56–60	1	2	3
11–15	165	168	333	61–65	2	—	2
16–20	100	117	217	66–70	—	2	2
21–25	65	72	137	71–75	1	1	2
26–30	40	57	97	76–80	—	1	1
31–35	22	35	57	81–85	1	—	1
36–40	6	14	20	86–90	—	1	1
41–45	10	9	19	91–95	1	1	2
46–50	5	7	12				
				Total	608	604	1212

TABLE G

Gerson, Opera. *Selected samples*

For details see text

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	30	29	59	61– 65	7	4	11
6–10	85	81	166	66– 70	3	5	8
11–15	108	115	223	71– 75	2	—	2
16–20	101	90	191	76– 80	2	2	4
21–25	68	78	146	81– 85	—	1	1
26–30	46	66	112	86– 90	—	2	2
31–35	53	45	98	91– 95	1	2	3
36–40	28	32	60	—	—	—	—
41–45	28	25	53	111–115	1	—	1
46–50	22	19	41	—	—	—	—
51–55	14	8	22	131–135	—	1	1
56–60	7	6	13				
				Total	606	611	1217

TABLE H

Gerson, Opera. *Random passages*

For details see text

No. of words	Sentences			No. of words	Sentences		
	A	B	Total		A	B	Total
1– 5	23	34	57	61– 65	6	5	11
6–10	99	97	196	66– 70	4	6	10
11–15	97	111	208	71– 75	3	2	5
16–20	105	98	203	76– 80	2	2	4
21–25	75	80	155	81– 85	1	1	2
26–30	48	53	101	86– 90	1	—	1
31–35	43	26	69	91– 95	1	1	2
36–40	32	33	65	96–100	1	—	1
41–45	25	16	41	—	—	—	—
46–50	19	20	39	121–125	1	—	1
51–55	6	9	15	126–130	1	—	1
56–60	7	6	13				
				Total	600	600	1200

TABLE J

Graunt's Observations upon the Bills of Mortality

A, B, C, first, second and third portions: the whole included apart from some omissions (see text)

No. of words	Sentences				No. of words	Sentences			
	A	B	C	Total		A	B	C	Total
1– 5	—	—	—	—	86– 90	2	4	2	8
6–10	3	2	7	12	91– 95	2	—	1	3
11–15	2	9	2	13	96–100	—	1	4	5
16–20	5	9	9	23	101–105	1	—	—	1
21–25	8	12	9	29	106–110	1	1	1	3
26–30	8	11	6	25	111–115	—	—	—	—
31–35	12	8	20	40	116–120	1	2	—	3
36–40	10	10	8	28	121–125	1	1	1	3
41–45	8	8	8	24	126–130	2	—	—	2
46–50	8	6	1	15	131–135	—	—	—	—
51–55	9	9	5	23	136–140	1	—	—	1
56–60	8	3	4	15	—	—	—	—	—
61–65	4	3	5	12	151–155	—	—	1	1
66–70	5	4	6	15	156–160	—	1	1	2
71–75	5	2	5	12	—	—	—	—	—
76–80	3	3	2	8	211–215	—	1	—	1
81–85	2	2	4	8					
					Total	111	112	112	335

TABLE K

Petty

A, *Political Arithmetic*, 300 sentences, with 34 added from the *Treatise of Taxes*. B, *Treatise of Taxes*. C, random passages (see text)

No. of words	Sentences			No. of words	Sentences		
	A	B	C		A	B	C
1– 5	1	1	—	131–135	5	1	2
6– 10	4	3	6	136–140	3	2	2
11– 15	3	8	13	141–145	2	4	3
16– 20	11	21	17	146–150	4	3	4
21– 25	16	17	26	151–155	2	1	2
26– 30	20	20	31	156–160	1	—	1
31– 35	26	16	30	161–165	1	3	—
36– 40	22	31	27	166–170	1	1	—
41– 45	18	28	24	171–175	1	1	—
46– 50	28	19	18	176–180	—	—	—
51– 55	12	18	11	181–185	1	—	—
56– 60	21	15	14	186–190	—	1	—
61– 65	23	14	16	191–195	—	—	—
66– 70	16	16	11	196–200	1	—	—
71– 75	10	13	10	201–205	—	—	—
76– 80	14	15	8	206–210	—	—	1
81– 85	10	7	12	211–215	2	1	2
86– 90	14	10	11	216–220	—	—	—
91– 95	6	9	5	221–225	1	—	1
96–100	5	8	4	226–230	—	—	—
101–105	3	4	2	231–235	1	—	1
106–110	5	10	5	236–240	—	—	—
111–115	3	2	1	241–245	1	—	—
116–120	4	8	5	—	—	—	—
121–125	5	3	5	256–260	1	—	—
126–130	6	—	3				
				Total	334	334	334