Cambridge University Press
978-1-107-63371-1 - The Statistical Study of Literary Vocabulary
G. Udny Yule
Excerpt
More information

Chapter 1

INTRODUCTORY, PERSONAL AND APOLOGETIC

1.1. Words are to the writer what paints are to the painter, the materials at his disposal for the purpose of creation. Carefully selecting, arranging and re-arranging words from his treasury, he shapes a poem or a tale for our delight, a history or some grave tractate for our edification. To those who write, many of us owe more than to almost any other class of men. How kind, how more than kind, they have been thus to create for our pleasure, and what a magical work they have done. The Word, by a sacred metonymy, symbolises the eternal spirit of Creation: the word *spoken* by the lips of humanity is a very emblem of the fleeting, gone down the wind with the breath that uttered it—

Nam et ipsi qui loquuntur,
ecce omnes nihil;
deficient enim cum sonitu verborum.

Thee too the years shall cover; thou shalt be
As the rose born of one same blood with thee,
As a song sung, as a word said.

But the writer takes these emblems of the fleeting and endows them with life, a life it may be outlasting many generations of men, a life that on rare and happy occasions audaciously we term immortal.

1.2. To pin down such creations and dissect them is but a sorry business, work for the second-rate, a job for the uncreative. For the uncreative can at least ask questions, and very foolish questions they may ask at times: Did Thomas à Kempis really write that little volume which passes under the title of its first chapter, the *De Imitatione Christi*?* Did Shakespeare write the plays that are generally attributed to him, or such and such a particular play? Did St Paul write the Epistle to the Ephesians? What is the probable chronological order of this, that and the other work of Plato? In endeavouring to obtain an answer to any such question every element of that highly complex quality the author's style may and should be taken into account, but amongst those elements his *vocabulary*—the aggregate of words he uses—takes an important position. It is a definite characteristic, and

* 'Non quaeras quis hoc dixerit sed quid dicatur attende.' *Imit.* I. 5.

certain aspects of it at least are susceptible of definite numerical treatment, treatment by what we now term statistical methods.

1.3. It was no mere desire to get another field in which to exercise my craft that led me to take up the present work. Interest in the *De Imitatione Christi* provided the initial stimulus. The controversy on the authorship of that book seemed to me mostly quite foolish, a matter of historical and pathological rather than actual interest, but I had read some few of the works relating thereto and in these, from comparatively early days, the vocabulary and diction of Thomas à Kempis are discussed as evidence. These discussions left in my mind a sense of inadequacy. They did not tell me what *I* wanted to know. They dealt with such details as his use of words and idioms taken literally from the Dutch—like the well-known *exterius* in the sense of *by heart* in Lib. I, cap. i of the *Imitatio*; of words used in unusual, non-classical senses; of italianate words, and so forth. All these are mere details, details certainly quite useful in relation to the controversy, providing valid evidence (if they are rare elsewhere) when we find them occurring both in the *Imitatio* and in other works admittedly by Thomas; *but they give no faintest notion as to what his vocabulary is really like as a whole*. To tell me that there is a small mole on Miranda's cheek may help me to identify the lady, and may in conceivable circumstances be quite useful information to the police, but it hardly amounts to a description of her alluring features.

1.4. What I felt I wanted in the first place, prior to any detail, was some summary, some picture of the vocabulary *as a whole*. Surely the colour and flavour of a text, if I may be permitted to mix my metaphors, are determined not by the exceptional words, unless these words taken together form a large class, but in the main by the common words used by the author, the words used by him over and over again? This was the question I asked myself. The sort of picture I wanted could only be given by a list of the words used classified by the number of times they were used; words used once, words used twice, words used thrice and so on.

1.5. I am no linguist: I was wholly ignorant of what had been done in the way of applying statistical methods to the study of vocabulary, and know very little of that 'literature' now. Rashly no doubt in the circumstances, but purely for my own interest, I started an exploratory investigation on the lines suggested, an investigation of which full details are given in Chapters 9 and 10. Naturally enough the investigation was devoted to the work which had excited my interest in the question, the *Imitatio*: it is a conveniently short work, of some 42,000 words only, and the fact that there is a complete concordance available is a further convenience. To limit material, I decided to confine myself to a single class of words, viz. nouns. The concordance was worked through page by page and every noun entered on a card together with the number of times it was used. From these cards it was easy to book up a table, the 'frequency distribution' to use the statistical term, showing the numbers of nouns used once, twice, thrice, etc., and also a list showing the actual nouns used once, twice, thrice, etc. This second classification was what I set out to obtain, and it presented many points of interest. But one never knows what one is going to find in the course of such a bit of exploration, and quite unexpectedly the simple frequency distribution showing merely the *numbers* of nouns used once, twice, thrice, etc. proved to have considerable interest of its own. It showed that little less than half—some 45 per cent—of the nouns were used only once in the whole work, some 15 per cent twice, 9·5 per cent thrice, and so on in a very long gradually and irregularly decreasing series, only terminating with a word which occurred 418 times—very appropriately the noun *Deus*. The large proportion of nouns used only once was a complete surprise, and the whole form of the distribution at once raised the question how far peculiarities of form might be characteristic of one author as compared with another.

1.6. The interest of the results was quite sufficient to encourage me to proceed, and I decided to draw up comparative vocabularies for (1) the miscellaneous works of Thomas à Kempis other than the *Imitatio*, (2) the works of Gerson—one of the competitors, with many supporters in former days and even in the latter part of the last century, for the authorship of the *Imitatio*. But here there could be no question of completeness. Even if there had been concordances available to the other works of Thomas and to the works of Gerson they would have been of little use, for the material would have been unwieldy, and the great length of the works as compared with the little *Imitatio* would have rendered many comparisons difficult or almost impossible. It was decided accordingly to proceed by way of sample and, for convenience of comparison and avoidance of possible fallacies, to make the sample taken from each of our authors of about the same size as the *Imitatio*, that is to say a sample of rather more than 8000 occurrences of a noun. Passages were taken well spread over the miscellaneous works of Thomas and the theological works of Gerson, and the occurrences of nouns booked up on to cards, using the same cards as for the *Imitatio* in so far as the nouns were the same. When the work was done, every card had an entry for (1) the number of times the noun occurred in the *Imitatio*, (2) the number of times it occurred in the sample from the miscellaneous works of Thomas, (3) the number of times it occurred in the sample from the theological works of Gerson. The results were summarised in the same way as before, and of course there were now many further points of interest, since not only were comparisons possible, but other statistical methods could be used. The two new frequency distributions resembled in general form that for the *Imitatio*, but were more contracted, not extending into so long a 'tail' of words used many times: *Deus* remained the substantive most frequently employed, but it occurred only 369 times in the sample from the miscellaneous works of Thomas and 256 times in the sample from Gerson, as against 418 times in the *Imitatio*.

1.7. But all this work had raised one question after another of a purely statistical kind, questions some of which, so far as I know, had never either been asked or answered. What is the nature of these frequency distributions showing the numbers of nouns used once, twice, thrice, etc.? It is evident that if we go on extending the size of sample taken from a given work, or a given author's 'works', the distribution will itself continually extend. If in 2000 occurrences of a noun taken at random from the given work the most frequent noun occurs some 50 times, in 4000 occurrences of a noun taken at random from the same work it will probably occur some 100 times, more or less. Everything about the distribution—or apparently everything—will in fact alter as the size of sample is increased: the mean number of occurrences of a noun will go up; its reciprocal, the number of distinct or different nouns per 1000 occurrences of a noun, will go down; any measure of scatter or dispersion will rise rapidly; the percentage of nouns used only once will fall slowly. Are not some of the statistical measures or percentages that have been used in practice fallacious for these reasons? Can one throw any further light on such fallacies, or possible sources of fallacy? Is there no characteristic of the distribution that is independent of size of sample?

1.8. In quite the early days of the work, the analogy had forced itself on my mind between these data showing the numbers of words used once, twice, thrice, etc. and data showing the numbers of persons out of a given number at risk who have met with 0, 1, 2, 3, ... accidents during a given period of exposure. Both are data respecting 'repeated events' or 'multiple happenings' as they have been termed. But there are two important points in which the verbal case differs from the case of personal accidents. In the first place, the case of personal accidents involves time: it presents time-problems. Given say the numbers of persons out of 2000 at risk who have met with 0, 1, 2, 3, ... accidents during a year, we can ask ourselves for example how many will have met with 0, 1, 2, 3, ... accidents during a longer period of exposure, 2 years or 5 years. Time in the personal case is replaced in the verbal case by the total number of occurrences of a word (or noun or whatever it may be) on which we have based our table, i.e. the amount of material, roughly the number of pages of the author, that we have used. We are not in the least concerned to know how long the author took to write those pages. And it may be noted that this change of aspect almost inevitably leads to a certain change of nomenclature. It is hardly possible to avoid terming the amount of material on which we base our table, viz. the total number of occurrences of a word, the *size of sample,* but in doing so we are using the term in a sense wholly different from that which it would normally carry in the case of personal accidents, viz. number of persons at risk. Our 'amount of material' or 'size of sample' is most nearly equivalent to 'time of exposure to risk' in the accident case, for the total number of accidents is, within the limits of fluctuations of sampling, directly proportional to time. This first point of difference is an essential, but not a very troublesome distinction. In the second place, while in the case of personal accidents we know the number of persons at risk and consequently the number who have not met with any accident during the period of exposure, we do *not* know the number of words 'at risk', that might have been used by the author, and cannot insert in our table the leading figure 'words occurring 0 times'. We may be able to make a very rough guess whether the total vocabulary is nearer 5000, or 10,000, or 20,000, or 50,000, but that is about all. Our table for words is, therefore, compared with its analogue, incomplete, lacking its head, decapitated. This is not only an essential but an emphatically troublesome distinction.

1.9. The theory of accident-distributions has received a great deal of attention during the past twenty-five years or so, and that theory was familiar to me. A certain characteristic of the distribution which, in greatly simplified hypothetical circumstances not strictly valid for any real case, is independent of the period of exposure to risk, was known. It proved possible to transform this characteristic into a form applicable to the decapitated table for words. One had therefore now got that obvious desideratum, a figure characterising the word-distribution which was independent of size of sample within the limits of mere fluctuations of sampling. This seemed an important result, so important that.it was worth an experimental test. The test was accordingly carried out, by taking a series of samples spread over one and the same work, and gave satisfactory verification. The 'characteristic' remained the same within the limits of fluctuations of sampling whether the distribution was based on some 2000, 4000, 6000 or 8000 occurrences, and one was therefore able for the first time to compare two distributions without regard to the numbers of occurrences on which they were respectively based. Full details of the experiment are given in Chapter 4. The results gave one confidence in the general notions on which the theory was based, and led one to apply them to the consideration of possible fallacies (Chapter 5). In this work the experimental data again came in useful as illustrations. Next, the 'characteristic' having been obtained, it was obviously desirable to find out the extent to which it would be likely to vary in data drawn from different but similar works of one and the same author. This information could only be got by actual trial, and the trial was duly carried out, using four of Macaulay's essays as one example and four of Bunyan's works as another (Chapter 6). This again raised further questions. The data provided by the work on Macaulay and Bunyan showed certain striking similarities as regards the relative numbers of words peculiar to each separate work or common to any specified two, three, or all four of the works. To what are these similarities to be ascribed? The answer to the problem, discussed in Chapter 7, is that they depend simply on the form of the word-distribution—the proportionate numbers of words that are used by the author once, twice, thrice, etc. Again, when the first card drawers for the Bunyan data and for the Macaulay data happened one day to be standing open together, it leapt to the eye that while in Macaulay nouns beginning with A were much more numerous than those beginning with B, in Bunyan matters were exactly reversed, nouns beginning with B being by far the more numerous. Other more or less conspicuous differences between the alphabetical distributions were then noticed, and this led finally to a long investigation (Chapter 8), including an analysis of the respective vocabularies into words of Old English-Teutonic and of Latin-Romance origin. The difference between the two alphabetical distributions is mainly due to the much larger proportion of Latin-Romance nouns in Macaulay.

1.10. It is evident from this very summary account that my work proceeded very much as such a piece of research work is apt to proceed, opening up problems that were quite unforeseen at the start and inevitably changing, to a greater or less degree, one's point of view. As the work advanced my original notion of simply trying to obtain a picture of the vocabulary by means of the classified word list, though remaining important, faded in some degree into the background, and purely statistical problems took its place—for their solution was essential to the understanding of what one was doing and to the interpretation of results. Indeed I might add their solution was essential to the understanding of what *other* people had been doing and to the interpretation of *their* results. There remain many bits of investigation which I should like to undertake to fill in obvious gaps and illuminate obscure points, but work of this kind takes a long time and they might occupy years. Years are fleeting, and I am old. I have thought it best to give some account of results reached at the present stage, hoping that, however incomplete, they may at least be suggestive, and that errors may be corrected, omissions supplied and further advances be made by others. This is no logically ordered text-book, but a collection of notes. All that I have been able to do in the way of logical ordering is to give an exposition of more theoretical and general notions first, and postpone to the end the investigation into the vocabularies of Thomas à Kempis and Gerson which actually initiated my work. Publication of that investigation does, I know, require apology, and I have already apologised, for I am no linguist. But after all neither the statistician without training in linguistics, nor the linguist without training in statistical method, is properly equipped for work in this field, and the latter has charged in quite cheerfully, not always without disaster.

1.11. It seems almost strange that no statistician, so far as I am aware, should have specially devoted himself to this branch of work during the past half-century of rapid advance in statistical method. It is full of interest, and is of the most venerable antiquity and august associations, for the earliest examples that one can give relate to the Hebrew text of the Old Testament. To cite first from Dr Wheeler Robinson (see References at the end of the chapter):

> After the destruction of the Jewish state in A.D. 70, it obviously became necessary to ensure the preservation of this consonantal text [of the Bible] and the correct tradition of its pronunciation. This became ultimately the work of the *Masoretes*. The name 'Masoretes' is derived from the Hebrew word *Masora*, in the sense of 'tradition'.
>
> ...in general they were meticulously careful guardians of the consonantal text *as it reached them*. This is seen from the general character of the Masora, i.e. the textual notes which are found written above, below, and in the margins of Biblical manuscripts, as well as in separate treatises. These are largely concerned with the indications of *hapax legomena*, or of the number of instances in which a particular form occurs. Everything countable seems to be counted, and if such work seems to us often futile, it was not so to men deeply concerned with the literal accuracy and the 'plenary inspiration' of the Scriptures; moreover, its practical usefulness for copyists is apparent.

The writer in Hastings's *Dictionary of the Bible* somewhat expands this statement as to the work of the Masoretes:

> They *counted* the verses and the words of each of the 24 books and of many sections; they reckoned which was the middle verse and the middle word of each book; nay, they counted the letters both of particular sections and even of whole books.

Thus they could specify the middle word and the middle letter in the Torah, the middle verse and the middle letter of the Psalms:

> They counted also the frequency of occurrence of words, phrases, or forms, both in the whole Bible and in parts of it.
>
> They collected notabilia into groups, and thus not only helped the recollection of these, but also facilitated the control of the MSS.... There is a great fondness for anything alphabetical; e.g. we have an alphabetical list of words which occur only twice in the O.T.—once with and once without [the conjunction] *waw* at the beginning.

Much of this, it must be confessed, does seem 'futile', and its 'practical usefulness for copyists' is not very apparent. Surely counting the number of times a given word occurs (by no means an easy task to perform with accuracy) or the number of letters is a strangely indirect and insufficient way of checking the accuracy of a copy? The whole process seems odd and puzzling, but a passage in the *Jewish Encyclopaedia* makes a very plausible suggestion as to its origin:

> In classical antiquity copyists were paid for their work according to the number of stichs. As the prose books of the bible were hardly ever written in stichs, the copyists, in order to estimate the amount of work, had to count the letters. Hence developed, in the course of time the Numerical Massorah, which counts and groups together the various elements of the text.

This suggestion also accounts in a very natural way for the practice of determining the middle letter, or word, or verse of a book, which otherwise seems quite pointless. The tired copyist would always be glad to know when he had completed half of his task.

1.12. The average non-mathematical student of literature will, I am afraid, find much of this work very difficult. I have endeavoured to help him, so far as may be, by explanations in general terms, by practical examples, and by giving at the end of each chapter, where it is necessary, references to a text-book on statistical method. Regarding the numeration of references, paragraphs, etc. the reader is referred to the note on p. viii.

REFERENCES

My citations in § 1.11 concerning the Masora and Masoretes are made from:

(1.1) Wheeler Robinson, H., Editor (1940). *The Bible in its Ancient and English Versions.* Oxford: Clarendon Press. (See pp. 26, 29.)

(1.2) Hastings, J., Editor (1898–1904). *The Dictionary of the Bible.* Edinburgh: T. & T. Clark. (See the article 'Text of the Old Testament', especially Section iii, The work of the Mas(s)oretes.)

(1.3) Singer, I., Editor (1901–6). *The Jewish Encyclopaedia.* New York and London: Funk & Wagnalls. (See the article 'Masorah'.)

Cambridge University Press
978-1-107-63371-1 - The Statistical Study of Literary Vocabulary
G. Udny Yule
Excerpt
More information

Chapter 2

THE WORD-DISTRIBUTION: ILLUSTRATIONS

2.1. The reader will have gathered from Chapter 1 that in the present work a position of importance is assigned to the table showing the number of words used once, twice, thrice, etc. by an author in any given writing or sample from his writings. The 'words' in question may include all words, or nouns or adjectives or verbs only, or those words only that fall within some special class defined by the investigator. Let us look at some specimens of such tables.

2.2. *Illustration* 2.1. Columns 1 and 2 of Table 2.1 give the data for nouns, Latin nouns, in the *De Imitatione Christi*. The table was compiled from Storr's *Concordance* (ref. 2.2): fuller details, references and notes on the practical difficulties encountered will be found in Chapter 9. It will be seen that on my reckoning there are 520 nouns that occur once only in the whole work, 174 that occur twice, 111 that occur thrice, and so on. The figures tail away very slowly, naturally with slight irregularities here and there, and the table only terminates with a noun that occurs 418 times, the noun *Deus* as already mentioned in Chapter 1. In all there are 1168 separate and distinct nouns in the table (see figure at the foot of col. 2), or a *vocabulary* of 1168 nouns to put it briefly. The number f_x of nouns that occur X times is called in statistical terms the *frequency* of X, and the manner in which the frequencies are distributed over the scale of X is spoken of as the *frequency distribution* of X. In form this frequency distribution shows a maximum at the bottom end of the range (the top of the table) for words occurring once only, and from this point the frequencies decrease at first rapidly and then more and more slowly.

2.3. Let us study this first illustration in some detail. In col. 3 of the table are given the products $f_x X$: each such product evidently gives the number of occurrences of a noun due to nouns of the class X (nouns that occur X times). The total number of occurrences, as given at the foot of the table, is 8225. Col. 4 gives the sum of the figures in col. 2 added up step by step from the bottom, so that the figure on line X gives the number of nouns occurring X times *or more*. Col. 5 similarly gives the sum of the figures in col. 3 added up step by step from the bottom, so that the figure on line X gives the number of occurrences due to nouns occurring X times *or more*.

TABLE 2.1. Showing the number f_x (col. 2) of nouns occurring X times (col. 1) in the *De Imitatione Christi*: compiled from Storr's *Concordance* (ref. 2.2). For explanation of cols. 3–7 and data at foot see text. Note that to save space blank rows are omitted in this and all similar tables

1	2	3	4	5	6	7
X	f_x	$f_x X$	$S(f_x)$ from bottom	$S(f_x X)$ from bottom	Col. 4 reduced to total 10,000	Col. 5 reduced to total 10,000
1	520	520	1,168	8,225	10,000	10,000
2	174	348	648	7,705	5,548	9,368
3	111	333	474	7,357	4,058	8,945
4	70	280	363	7,024	3,108	8,540
5	37	185	293	6,744	2,509	8,199
6	33	198	256	6,559	2,192	7,974
7	20	140	223	6,361	1,909	7,734
8	28	224	203	6,221	1,738	7,564
9	11	99	175	5,997	1,498	7,291
10	14	140	164	5,898	1,404	7,171
11	10	110	150	5,758	1,284	7,001
12	9	108	140	5,648	1,199	6,867
13	11	143	131	5,540	1,122	6,736
14	5	70	120	5,397	1,027	6,562
15	4	60	115	5,327	985	6,477
16	7	112	111	5,267	950	6,404
17	7	119	104	5,155	890	6,267
18	4	72	97	5,036	830	6,123
19	5	95	93	4,964	796	6,035
20	2	40	88	4,869	753	5,920
21	5	105	86	4,829	736	5,871
22	1	22	81	4,724	693	5,743
23	1	23	80	4,702	685	5,717
24	7	168	79	4,679	676	5,689
25	2	50	72	4,511	616	5,484
26	1	26	70	4,461	599	5,424
27	4	108	69	4,435	591	5,392
28	3	84	65	4,327	557	5,261
29	2	58	62	4,243	531	5,159
30	3	90	60	4,185	514	5,088
31	1	31	57	4,095	488	4,979
33	2	66	56	4,064	479	4,941
34	2	68	54	3,998	462	4,861
36	2	72	52	3,930	445	4,778
37	5	185	50	3,858	428	4,691
38	4	152	45	3,673	385	4,466
39	3	117	41	3,521	351	4,281
40	1	40	38	3,404	325	4,139
41	1	41	37	3,364	317	4,090
43	2	86	36	3,323	308	4,040
44	2	88	34	3,237	291	3,936
46	2	92	32	3,149	274	3,829
48	1	48	30	3,057	257	3,717
50	1	50	29	3,009	248	3,658