THE STATISTICAL STUDY OF LITERARY VOCABULARY

THE STATISTICAL STUDY OF LITERARY VOCABULARY

BY

G. UDNY YULE, C.B.E., M.A., F.R.S.

FELLOW OF ST JOHN'S COLLEGE, AND
FORMERLY READER IN STATISTICS,
CAMBRIDGE

CAMBRIDGE
AT THE UNIVERSITY PRESS
1944

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Published in the United States of America by Cambridge University Press, New York

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107633711

First published 1944
First paperback edition 2014

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-63371-1 Paperback

CONTENTS

PREFACE page ix

CHAPTER 1. Introductory, Personal and Apologetic 1

1.1–1.3. The inception of this work was due to interest in the vocabulary of the *De Imitatione Christi*: previous work seemed too much concerned with details. 1.4. I wanted a general picture, which would only be obtained by tabulating separately words used once, twice, thrice, etc. 1.5. The work was carried out, with a concordance, for nouns only, and 1.6, results were sufficiently interesting to lead to further work, by sampling, on the miscellaneous works of Thomas à Kempis and the religious works of Gerson. 1.7–1.9. But this work showed how ignorant one was of the nature of such statistics: the special investigation was laid aside and general investigations undertaken to solve problems essential to the understanding of what one was doing. 1.10. Thus the bulk of this work is devoted to problems quite unforeseen at the start, and the original investigation only takes a minor place as an illustration at the end. 1.11. Statistics of literary vocabulary are of a respectable antiquity, as instanced by the statistics of the Masoretes. 1.12. A note on references. References. *Pages* 1–8.

CHAPTER 2. The Word-distribution: Illustrations 9

2.1. The word-distribution, showing how many words are used once, twice, thrice, etc. is important in this work: 'words' may include all words or any special class. 2.2–2.6. Illustration 2.1. Nouns in the *De Imitatione Christi*: detailed analysis: mean, standard deviation etc. 2.7. Illustration 2.2. Nouns in samples from the works of Thomas à Kempis. 2.8. Illustration 2.3. Nouns in samples from religious works of Gerson. 2.9. Illustration 2.4. Nouns in Macaulay's essay on Bacon. 2.10. Illustration 2.5. A limited class: non-Vulgate nouns in the sample of Illustration 2.3. 2.11. Illustration 2.6. Adjectives in the *Imitatio*. 2.12. Illustration 2.7. Verbs in the *Imitatio*. 2.13. Was the choice of nouns for the first investigations a wise one? Comparative neutrality of adjectives and verbs, especially the latter, as regards subject. 2.14–2.15. Illustration 2.8. Approximate distribution for nouns in the Vulgate, as an example based on a much larger sample. 2.16–2.18. Illustration 2.9. Nouns in St John's Gospel in Basic English, illustrating effect of a limited vocabulary and, 2.19, Illustration 2.10, in St John's Gospel, A.V., for comparison. 2.20–2.23. Some general deductions from the tables. 2.24. Terminology: 'vocabulary' and 'occurrences'. 2.25. The tables of Zipf are not comparable with those given, since he enters separately every inflection of a word etc. References. *Pages* 9–34.

CHAPTER 3. Theory of the Word-distribution, with an Exordium on Sampling: the Characteristic 35

3.1–3.5. General notions on sampling, practical methods of random sampling, the standard error of a mean. 3.6. Difficulties of random sampling for vocabulary. 3.7–3.9. Spread sampling as an alternative: advantages and disadvantages. 3.10–3.11. The word-distribution is a distribution of 'multiple happenings' or 'accidents', but 3.12, differs from a distribution of personal accidents in as much as (1) time is not specifically involved, (2) the distribution is decapitated, the number of words that have not met with the accident of being used being unknown: 'size of sample' in the personal-accident case means number of persons at risk: the number of words at risk is unknown, and 'size of sample' inevitably comes to mean the total number of occurrences, an equivalent of total number of accidents or of time. 3.13. For the accident-distribution we know a function of mean and variance which is (in the ideal case) independent of time, measuring the variance of members of the group for liability to accident. 3.14. Words also vary in liability to the accident of being used, and, 3.15, a quantity K derived from the

expression for accidents is a characteristic of the decapitated word-distribution which is independent of size of sample: but, 3.16, constancy will not be exhibited if the sample is extended, not by further random sampling from the same field, but by taking successive contributions to the sample from different portions of the author's works. 3.17–3.24. More complete development and discussion of the preceding general notions. 3.25. Possibility of characteristics of a higher order. 3.26. The problem of standard errors in relation to word-distributions. 3.27. An equation that fitted well some distributions for personal accidents failed, so far as tested, with word-distributions: a formula given by Zipf also fails. References. *Pages* 35–56.

CHAPTER 4. The Characteristic K: Practical Problems 57

4.1. Desirability of a practical test of the constancy of the characteristic for different sizes of sample: values of the characteristic for distributions of Chapter 2. 4.2–4.12. An experiment in sampling, testing the constancy of the characteristic for samples of 2000 to 8000 occurrences of a noun from Macaulay's essay on Bacon. 4.13–4.23. Attempts at estimating the order of magnitude of W, the total number of words at risk: further discussion of the meaning of the term. 4.24. Hence estimates of the value of v_λ. 4.25–4.31. An empirical relation between characteristic and mean for samples of the same size: discussion. 4.32. The range of variation of K to be expected in different but similar works of the same author is an important practical question: see Chapter 6. References. *Pages* 57–82.

CHAPTER 5. Certain Difficulties and Sources of Fallacy 83

5.1–5.13. The ratio of the vocabulary of author A to that of author B in samples of the same size is a function of size of sample. For a small sample the limit of the ratio is unity in all cases, for a large sample W_2/W_1. The course of the ratio between these limits may be very varied. Difficulty of interpreting an impressionist judgment. 5.14–5.23. Proportions or percentages. Two forms of statement are in use for assessing the proportion of some class of words to the whole: (1) the proportion p_w of occurrences of words of the special class to all occurrences, (2) the proportion p_v of words in the special list to the total of words in the vocabulary. The first, p_w, is independent of size of sample: the second, p_v, is a function of size of sample. 5.17. Illustration 5.1. Nouns not in Lewis and Short's *Dictionary* in the data from the *Imitatio*, etc. 5.18. Illustration 5.2. Once-nouns in Macaulay's essay on Bacon. 5.19–5.23. Illustration 5.3. Romance words in Chaucer. 5.24–5.26. The proportion, p_c, of special vocabulary to total occurrences or some statistical measure thereof, such as lines or pages, tends continuously to decrease with increasing size of sample. 5.27–5.28. Applicability of these conclusions in fields other than vocabulary. 5.29. The reason for this exceptional trouble, the various proportions being functions of size of sample, is that we are not using that expression in its usual sense—but such usage seems quite unavoidable: Lutoslawski over forty years ago wrote of 'samples' of text. 5.30. Note on a paper by G. H. Thomson and J. R. Thompson. References. *Pages* 83–116.

CHAPTER 6. Word-distributions from different Works of the same Author 117

6.1. The desirability of tests to see how far word-distributions based on different but similar works of the same author are consistent with one another. 6.2–6.6. First test: data for three additional essays of Macaulay, making four in all. 6.7. Distribution for the pool of the four essays: the characteristic reduced. 6.8–6.13. Second test: data for four of Bunyan's works. 6.14. Distribution for the pool of the four works: the characteristic reduced. 6.15 *et seq.* On certain classes of nouns in Bunyan and Macaulay: 6.16–6.18, Verbal nouns in *-ing*; 6.19–6.20, Monosyllabic nouns; 6.21–6.24, Biblical nouns. 6.25–6.26. Conclusion. References. *Pages* 117–147

CHAPTER 7. The Distribution of the Vocabulary over Samples from several Works of the same Author: Vocabulary Ratios of Bunyan to Macaulay 148

7.1–7.3. The distribution of vocabulary over four samples from the same author, i.e. over the classes *ABCD* (words in all four samples), *ABCδ* (words in A, B, C but not D), *ABγδ*, *Aβγδ*, etc.: the remarkable similarity in three different instances. 7.4–7.7. The problem of finding the distribution if the total vocabulary is partitioned at random into four (or generally *r*) equal samples. 7.8–7.10. Comparison with the observed data: general consilience, but with certain points of difference arising from heterogeneity of samples. 7.11. Possible measures of heterogeneity. 7.12. Summary of preceding paragraphs. 7.13–7.18. Associations between presence of a noun in one sample and in another from the same author: low values obtained: comparison of results of random partitioning with observation. 7.19–7.25. Vocabulary ratios of Bunyan to Macaulay: 7.20–7.22, a fallacious method of deducing from the data, and, 7.23, a method of deducing by calculation from the respective total word-distributions; 7.24, Source of the fallacy of §§ 7.20–7.22; 7.25, The method of § 7.23 seems the most hopeful yet suggested for determining the course of the vocabulary ratios with size of sample. References. *Pages* 148–182.

CHAPTER 8. The Alphabetical Distribution of English Vocabulary: Etymological Analysis of the Bunyan and Macaulay data 183

8.1–8.3. The remarkable difference between the alphabetical distributions of nouns from Bunyan and nouns from Macaulay while, 8.4–8.9, each of these authors is very self-consistent. 8.10. Probability that the difference is due to a larger proportion of Latin-Romance nouns in Macaulay. 8.11–8.14. Analysis of the alphabetical distributions into I. OE.-Teutonic and II. Latin-Romance nouns: the difference between them is mainly due to the greater proportion of Class II nouns in Macaulay. 8.15–8.17. The principal distinguishing characteristics of the distributions for Class I and Class II, and their possible uses for diagnosis. 8.18. Comparison of the Class II distribution with a distribution for Latin nouns. 8.19. Some further comments on the alphabetical distributions. 8.20–8.25. Etymological analysis of the word-distributions for the two authors: 8.20–8.22, All nouns; 8.23–8.25, Monosyllabic nouns. References. *Pages* 183–220.

CHAPTER 9. The *De Imitatione Christi*, Thomas à Kempis and Gerson 221

9.1. Origination of the work from interest in the *De Imitatione Christi* and the authorship controversy. 9.2–9.5. The first work on the nouns of the *Imitatio* itself, based on a concordance. 9.6. Further work, on the minor writings of Thomas à Kempis and the religious writings of Gerson, necessarily implied sampling. 9.7–9.9. The work on Thomas à Kempis: details of procedure. 9.10–9.12. The work on Gerson: details of procedure. 9.13. Statistical discussion of the three word-distributions. 9.14. Marking of the cards for presence of the noun in, or absence from, the Vulgate and each of certain dictionaries. 9.15. Nouns not in the Vulgate: distributions in the three authors. 9.16. Nouns not in Lewis and Short's *Dictionary*. 9.17. Nouns not in the Vulgate nor in any one of three dictionaries. 9.18. Non-religious, non-Vulgate, non-classical nouns borrowed from the Greek. 9.19. Summary of evidence to this point. 9.20–9.22. Critical discussion of the *Scutum Kempense* of Amort. References. *Pages* 221–250.

CHAPTER 10. The *De Imitatione Christi*, Thomas à Kempis and Gerson, continued 251

10.1–10.7. Correlations and contingency coefficients. 10.8–10.14. 'Frequent' nouns, occurring twenty times or more in the samples from the respective works: nouns characteristic of the *Imitatio* (or Thomas à Kempis) on the one hand and of Gerson on the other. 10.15–10.17. Association tables between the presence of a noun in the samples from one author and its presence in the samples from another. 10.18. Alphabetical distributions of nouns in the *Imitatio* and the sets of samples from Thomas à Kempis and from Gerson: similarity of the three distributions. 10.19. First supplementary investigation: the distributions of 'non-classical' nouns, not in Smith's *Smaller Latin-English Dictionary* (1933). 10.20–10.23. Second supplementary investigation: distributions and characteristics K for four of the minor works of Thomas à Kempis. 10.24. Consensus of the evidence from vocabulary with prior evidence obtained from sentence-length. 10.25. The argument that the *Imitatio* cannot be the work of Thomas, since it is so much better than anything known with certainty to be his. References. *Pages* 251–280.

CHAPTER 11. Valedictory 281

11.1. Incompleteness of the present work: need for more work on adjectives and verbs. 11.2. For safety the work on nouns required a sample of some 2000 occurrences of a noun or about 10,000 words, a larger sample than is always available: the use of adjectives and verbs, and possibly other words as well, may reduce this requirement, but the best rules for practice need investigation. 11.3–11.4. The methods given are methods for studying language-in-use, and it is hoped they will not be used solely or mainly for matters of controversy. Some further problems for investigation by the student of language or literature, and, 11.5, some problems for the theoretical statistician. 11.6. It is hoped that the methods and ideas suggested may prove fruitful. 11.7. Postscript: Theorem on the Characteristic. *Pages* 281–283.

Table of characteristics for nouns and chart 284–285

APPENDICES: Nouns in the samples grouped according to the numbers of occurrences

I. Bunyan: A. *Pilgrim's Progress*, Part I 286

II. Bunyan: B. *Pilgrim's Progress*, Part II 289

III. Bunyan: C. *Mr Badman* 292

IV. Bunyan: D. *Holy War* 295

Notes on certain words in the preceding Appendices 298

INDEX 299

NOTE ON THE NUMERATION OF PARAGRAPHS, EQUATIONS, ETC.

Each paragraph in the book is distinguished by a number consisting of the number of the chapter in which the paragraph occurs prefixed to the number of the paragraph in that chapter, and separated from it by a period, e.g. § 6.10 means the tenth paragraph Chapter 6. A similar system of numeration is used in all other instances: thus 'equation (3.9)' means the ninth (numbered) equation in Chapter 3, 'fig. 5.1' means the first figure in Chapter 5, 'ref. 3.4' means the fourth reference in the list of references at the end of Chapter 3

PREFACE

My first chapter is so largely of the nature of a preface that here I may be brief. This book arose from a desire to study a particular vocabulary in a case of disputed authorship. When I had advanced some way in that particular study, it became only too clear into how thorny a field of statistics I had strayed. Statistics of literary vocabulary proved to have their own special problems, their own peculiar difficulties and sources of fallacy, which no one apparently had made any attempt systematically to explore. The special study was accordingly laid aside for later use as little more than an illustration, and other investigations were taken up with the aim, not of throwing light on this or that problem of literature, but of illuminating the way in which statistics of the kind behaved and exemplifying methods which could be used in discussing them. As is inevitable in any such case, problem after problem arose that was not foreseen at the start. The book therefore was not and could not have been planned: if the reader find in it any logical development, it is simply the logic of the natural growth of the investigations.

I have found the work of such absorbing interest that the reader, I hope, may be interested too. If he find any novel ideas, I trust he will use them: if errors, let him amend them and judge mercifully. Neither old age nor the anxieties of war are favourable to continuity or clarity of thought.

In conclusion I have to thank the Controller of H.M. Stationery Office for permission to reproduce the data of Tables 3.1 and 3.2 from the Reports cited, Mr C. K. Ogden, Mr H. Sykes Davies and Mr G. Herdan for several references that I have used, and the Syndics of the Press for the spirit they have shown in risking the issue of such a book as this at such a time.

G. U. Y.

May 1943