AUTHORSHIP PROBLEMS AND THE COMPUTER

H. COPPENS - IDE
Assistente R.U.G.

I owe my interest in the above subject, viz. the usefulness of the computer for the comprehension and the solution of authorship problems, to Professor A. Gerlo, Chancellor of the University of Brussels, who, some three years ago, suggested as a topic for my doctoral dissertation a study of the Julius Exclusus.
This pamphlet was written after the death of Pope Julius II (21 February 1513) and purposes to be an amusing and intellectual satire on the life of Pope Julius II.
However, and importantly, it is also an indirect criticisism of the papacy and of the Church. Thus, the value of the dialogue is twofold : it is, on the one hand, a literary gem of brilliant wit and, on the other hand, an important document belonging to the pre-Reformation period. Its authorship, which has constituted a problem ever since the pamphlet got known, has as yet to be solved. Among the alleged authors of the pamphlet are, for instance, Erasmus, Ulrich von Hutten, Faustus Andrelinus, and Hieronymus Balbi. The arguments from a historical, religious, and literary point of view which have been advanced in favour of the authorship of these respective authors are often very interesting, very ingenious, but altogether not very convincing, because the religious importance and the underlying tendency of the pamphlet sometimes colour the argumentation of philologists anxious to ascertain authorship in accordance with their own political-religious-philosophical views. Therefore, an objective method for attributing authorship should be looked for. Consequently, a computer-assisted statistical analysis of style has been considered. As it is not customary to give a public talk, or to publish anything, on findings relating to a doctoral dissertation - I hope to get finished with my work by 1972 - I have thought it preferable here to give a brief survey of some recent authorship studies.

Since some ten years the computer has been used, in one way or other, for the solution of authorship problems. The reason for this is not far to seek : only the computer allows of dealing with vast quantities of material, which it was hitherto scarcely thought possible to handle. As Louis Tonko Milic has said : "Those who conducted the attribution studies (i.e., with the help of the computer) dealt in millions of words and have lived to tell about it". What is the normal way to proceed in the study of authorship problems by means of the computer ? The text, which presents an attribution problem - the text of, say, x - is punched completely, if it is not too long. Then there are the possible authors, samples of whose works are punched. ("Sampling" is an interesting problem, which has many facets : choice of samples, size of samples, number of samples, etc. This, however, is a subproblem upon which it will not be possible

here to dwell as it would take us too far).
There are three factors determining the choice of the works of the authors from which samples are to be taken : quantity, genre, and period.
First of all, quantity, which should be statistically sufficient. It is evident that it is impossible to derive any valid conclusions from a total population of, say, 20 sentences. As for genre and period, these are philological requirements. Again, the importance of genre is evident. If, for instance, I write a letter to my husband in the United States, I'll word it in a quite different way from a letter addressed to the Vice-Chancellor of the University of Ghent, and the style of the latter letter will in its turn be different from that of a scientific paper, which, again, will be quite different from that of a novel, which in its turn will be quite different from that of a poem. (This is an empirical truth, which has been tested experimentally).
Period, also, is important. Older studies have proved that a person writes quite differently according as he is, for instance, twenty, forty, sixty or eighty. Thus we compare the work of x with work written around the same year, belonging to the same genre, and being sufficiently long.

When the punching is finished, a very rigorous checking should be done, as any mistake left in the input, will inevitably turn up in the output, not once but in each different test. The much-awaited optical scanner would be a great help here : its use would result in a gain of time, a gain of money, and a greater degree of certainty.
The particular usefulness of the computer resides in :

1) the possibility of combining (criss-crossing) the criteria;
2) the application of statistical tests.

I would like to mention here the influence of the computer on stylistic studies. I will not exaggerate and go so far as to say that the results obtained have been achieved thanks only to the computer. It is, however, a fact that the computer was greatly instrumental in introducing and promulgating the quantitative analysis of style. There have, of course, been some older studies, laboriously made before the arrival of the computer, but these were exceptions and were not as inclusive and, consequently, as conclusive as the recent ones.
A subjectively phrased assertion about style can, and should indeed, be replaced by an objective formulation.
Several recent attibution studies could serve as illustrations of what I have just said in my introduction.
Of course, attribution studies made without the benefit of the computer will in principle not be considered in this colloquy.
Also, I shall not treat of authorship problems concerned with poetry, as factors such as metre and prosody are too important here (Cf. my introduction : in my dissertation I am particularly concerned with Latin prose).
Nor shall I discuss some very important general works or works of a more theoretical nature on literary statistics (1), this to avoid overlapping.
Also to be omitted from my survey are studies concerned only with one or other subproblem, as these would be of less interest on the present occasion.
Research carried out by mathematicians or statisticians will likewise not be considered here as this is not

always of an agreeable and charming nature and so as to avoid unnecessary complications and early drawbacks (ex. the monumental Mosteller-Wallace study (2)).
Language also is sometimes important for establishing a criterion, but not in every case; therefore - and this in spite of my being a Latinist -, I will first take a close look at Ellegard's study of the Junius Letters, entitled "A Statistical Method for Determining Authorship" (3). A statistical method is, of course (see above), nothing new. In "The Statistical Study of Literary Vocabulary", published as early as 1944, Yule (4) already established a connection between vocabulary and statistics. What is new, however, is the use of the computer. How does Ellegard proceed ? The starting point of his working hypothesis is that authorship can be deduced from a study of style. Ellegard considers the criterion of "the distinctiveness ratio", viz. the criterion of plus- and minus words (a plus word being a word used more frequently by a given author than by his contemporaries, a minus word being a word less frequently used or not at all). Ellegard has read the Junius-letters from beginning to end and has noted, and remembered, some 458 words-expressions-syntactical units, which he considers, in the light of his intimate knowledge of the period concerned, to be plus or minus words. He then looks for these in the work of Sir Francis, the alleged author, and in a hundred texts of 10,000 words each. All this tremendous amount of work is done manually, and here, owing to a circulus-viciosus-reasoning, Ellegard makes a regrettable mistake, precisely because he does not use the computer at the initial stage of his study.
I do feel bad about mentioning this, as I have a profound admiration for Ellegard's tremendous achievement, as also for his previous brilliant studies (5), and highly valuable insights and conclusions. In the course of his study Ellegard comes to the realization that many of his 458 words are not relevant as criteria, as they occur in nearly the same proportion in the contemporary pamphlet literature, and has to conclude that a second test will be necessary; again he looks for these 458 words in some 100,000 words taken from political writings of the same period (published in the "Advertiser"). The danger inherent in this method is manifest. The criteria should be objectively deduced through computer-scanning so that we are sure that the discriminators are objectively arrived at and there is nothing even remotely subjective about them. Now, a complete vocabulary is study material, and index, reverse index, frequency list and concordance constitute the basic material from which a computer-assisted stylistic study should start. It is with a feeling of regret that in Ellegard's conclusions to his work we come upon the following words, which could not be more true : "For future work in this field of enquiry it would (...) be very desirable to have a complete inventory of the usage both of different periods of time, and of different kinds of literature" (p. 78).

Ellegard resorts to the computer only in the final stage of his research, viz., for the study of the end-product, and for the statistical processing of his data. In the Appendix he gives his lists and calculations.

It is but fitting that a classical philologist should talk at some length about the interesting work of the Reverend A.Q. Morton and about the studies he made in collaboration with Levison, Winspear, and Wake (6).

Morton - who studied theology, classical philology, and mathematics - always studies Greek texts : Plato, Homer, the New Testament.
Convinced of the importance of his work, I went to see him in Culross and saw him at work in Edinburgh University. Morton consistently works with the computer, which he uses from the very beginning the very end. He systematically combines several criteria, as he regards this criss-crossing as a kind of autocontrol. For instance, he combines sentence-length, the patterning of the conjunction "kai" (= and), and the use of the particle "de". Sentence - length is in itself a rather tricky criterion, about which a good deal has been written, from - I think - the early study of Sherman (1888) (7) to the recent study by O'Donnell (1970) (8). Moreover, I, too, had to consider sentence-length as a criteria for my study of the Julius Exclusus. A great many problems are inherent in sentence-length, for instance "What is a sentence ? ", "What about punctuation ? ". Morton does not consider content-words - typical content-words are substantives - as they are too much subject-bound. He prefers to study high-frequency words such as some prepositions, some pronouns, and the verb einai (to be). The combination of these criteria gives a kind of profile which yields insight into authorship.

I should not conclude this very short survey without mentioning Milic's (9) method of quantitative stylistic analysis, which he has applied to the authorship of "A Letter to Young Poet". Milic also looks for criteria that are not subject-bound, as appears from his statement that "Most studies of style specify (...) that this or that peculiarity is limited to prose of a certain type or on a certain subject. It has seemed to me necessary to go beyond this, to find a description which was insensitive to such change" (p. 80). Therefore he considers the total vocabulary qua content of a writer as "likely to be useless as a criterion of identification". Thus, Milic pre-codes the text into grammatical classes and subclasses (he adapts the categories of Fries); this pre-codes material is punched and the computer can scan different combinations. Milic takes into consideration "verbals", "syntactical variety" (three-word patterns), and "introductory connectives" as the most reliable discrimators for assessing authorship (in English prose). Milic proves his point conclusively in the tests made in connection with "A Letter to a Young Poet". Therefore I think it rather odd and not very sensible of a critic who recently reviewed one of the latest quantitave style-studies (10) that he should have written : "Im ganzen erhält man den Eindruck, dass das Ergebnis der aufgewandten Mühe und Arbeit nicht entspricht und das geisteswissenschaftliche Probleme nicht auf mathematische Weise und mit Rechenmaschinen zu lösen sind. Die Möglichkeit erscheint nicht ausgeschlossen, dass man mit dieser Methode je nach Wahl des Materials zu unmoglichen oder absurden Resultaten gelangen kann". The reviewer in question should read the very perceptive discussion by Milic, who considers the reliability of so many criteria, and tests and counter tests them, with a view to deriving from them not a subjective feeling of reliability but an objective list of facts.

Extrait de la Revue (R.E.L.O.)
VII, 1 à 4, 1971. C.I.P.L. - Université de Liège - Tous droits réservés.

Projects concerned with attribution studies are regularly listed in *Revue, Calculi* (11), and in *Computers and the Humanities* (12), which has a section entitled "Directory of Scholars active". In the Directory there have recently been mentioned such projects as Attribution of Rabelais' Fifth Book (chief investigator : Gerald J. Brault, Pennsylvania State University), Cyril Tourneur and *The Revenger's Tragedy* and *The Atheist's Tragedy* (investigator : Michael G. Farringdon, Swansea); Author Discrimination in Diderot's *Encyclopédie* (investigator : R.L. Frautschi, University of North Carolina); Indentifying Authors of Unsigned Newspaper Editorials (investigator : Wayne Allen Danielson, School of Journalism, University of North Carolina); Shakespeare and *Pericles* and *Anthony and Cleopatra* (investigator : Tommy Ruth Waldo, University of Florida); Hebraic Texts of Disputed Authorship (investigator : Asha Kasher, Department of Mathematics, Bar- Ilan University, Israel). On this latter investigation there has recently been published a preliminary report specifically concerned with the Isaiah passage in the Old Testament (13).

Radday describes several tests with a high diacritical capacity, most of which have been done previously. One of these tests, however, is completely original and of particular interest. It subdivides "special vocabulary" into six semantic groups, viz. war, material civilization, family, cult, society, and mind-emotions. This test, then, is a simplified, or embryonic, version of content analysis. All nouns were classified into one of the above six groups. The relative percentage of occurrences was established, and a graphical presentation of the results was made; these were plotted relatively to the one considered as normal, which belongs, of course, to the undisputed Isaianic part. The diagrams are really striking polygons (the norm being six radii of a circle), but, says Radday, "the statistical significance of these differences was not calculated, the classification of nouns into six semantic groups was after all necessarily governed by my subjective judgment..." (p. 73). It is a pity that Radday should not yet have had at his disposal a standardized list, which would have imparted a quality of objectivity to his truly interesting test. The test, as it now stands, has something arbitrary about is. It could, however, be improved in the direction of greater objectivity (14).

Some relevant facts emerge from this short survey. If we examine the choice of the criteria, it becomes apparent that the analysis of style has been oriented mostly towards subproblems such as word-length, sentence-length, whether combined or not with differential sentence-length, and also the study of vocabulary. Now, the study of style is not at all synonymous with the analysis of vocabulary in whatever sense this may be taken; the analysis of vocabulary is only a constituent element - though an essential one - of stylistic analysis. But the reason for resorting to vocabulary (i.e. lexical and/or grammatical items) for resolving authorship problems is obvious : vocabulary is more conducive to quantitative formulation and can be submitted to a digital computer. The study of combinations of wordclasses, word-groups in two, three, or even more wordpatterns is consequently, an excellent means of penetrating to the style of a given author. It is necessary here to stress the fact that the statistical analysis of style is a young method, and that every researcher is contributing to its development in his own way : by his choice of criteria, by his method of processing the data. Up to now quite different aspects have been considered, quite different methods have been followed, to accomplish the same fundamental aim, viz. the solution of attribution problems. After having passed in review these various aspects and methods, let us now briefly recapitulate.

Lexical items with meaning, and even short lexical strings and short clauses, with a high and a low these were Ellegard's criteria. At the opposite pole from "contentwords" stand "non-content-words" with a high frequency, which are insensitive to the subject treated (Morton). The frequency of occurrence of grammatical categories and subcategories, and a subtle combination of grammatical classes lead to a conclusive proof of the complexity of the style of Jonathan Swift (Milic). An embryonic attempt at content-analysis in terms of semantic groups shows a large spectrum of possibilities (Radda). All these who regard quantitative analysis of style as arid, monolithic and insensitive, or, in short, as unphilological, are, then, proved wrong : the diversity of possibilities, which this method presents, is astonishing. Also, in my own work I hope to prove that a broad spectrum of diacritical indicators yields a clear profile of an author, which is unmistakably distinguishable from that of his contemporaries.

The aim of this colloquy was to show the usefulness of the computer in various fields relating to the human sciences. The computer is part and parcel of our time. It would be anachronistic not to use its possibilities in every possible field, or to remain blind to them. Therefore, after having highlighted some interesting aspects of research done on the subject of authorship attribution, let us wonder at the modern beauty of computer research, and marvel, in Aristotelian sense, at the machine which permits for text-analysis, text-understanding, and, last but not least, data processing.

(1) (E.g.) Gustav Herdan, *Language as Choice and Chance,* Groningen, Noordhoff, 1956, XIII-356 p.; *Type-Token Mathematics,* 's-Gravenhage, Mouton, 1960, 448 p.; *The Calculus of Linguistic Observation,* ibidem, 1962, 271 p.; *Quantitative Linguistics,* London, Butterworths, 1964, XVI-284 p.; *The Advanced Theory of Language as Choice and Chance,* Berlin-Heidelberg-New York, Springer Verlag, 1966, XIII-459 p.
Pierre Guiraud, *Problèmes et Méthodes de la statistique linguistique,* Dordrecht, Reidel Publishing Company, 1959, 146 p.
Charles Müller, *Initiation à la statistique linguistique,* Paris, Larousse, 1968, 246 p. (= *Langue et langage)*; and his studies about Corneille.

(2) Frederick Mosteller & David L. Wallace, *Inference and Disputed Authorship.* The Federalist, Reading, Ma., Addison-Wesley Publishing Company, Inc., 1964, XV-287 p.,tabl. (= *Addison-Wesley Series in Behavioral Science : Quantitative Method*). "The primary purpose of their study was an exploration of Bayes' Theorem and a comparison of its possibilities for practical application with the more usual methods of statistical discrimination". (O'Donnell, p. 25, Cfr. note 8). Thus their work is, in their own words, "... a case study of the use of Baeyesian and other methods of discrimination." (p. 2).
Dieter Wickmann (Dr. rer. nat.), *Eine mathematisch-statistische Methode zur Untersuchung der Verfasserfrage literarischer Texte, durchgeführt am Beispiel der "Nachtwachen von Bonaventera" mit Hilfe der Wortartübergänge,* Köln-Opladen, Westdeutscher Verlag, 1969, 78 p., tabl. (= *Forschungsberichte des Landes Nordrhein-Westfalen,* n. 2019).

(3) Alvar Ellegard, *A. Statistical Method for Determining Authorship.* The Junius Letters (1769-1772), Acta Universitatis Gothoburgensis, 1962, 115 p., tabl. (= *Gothenburg Studies in English,* n. 13).

(4) G. Udny Yule, *The Statistical Study of Literary Vocabulary,* Archon Books, 1968², IX-306 p., tabl.

(5) Alvar Ellegard, *Estimating Vocabulary Size,* in *Word,* XVI, 1960, pp. 219-244, tabl.

(6) A.Q. Morton, *The Authorship of Greek Prose,* in *The Journal of the Royal Statistical Society,* Series A (General), Vol. 128, Part. 2, 1965, pp. 169-233, tabl.
A.Q. Morton, *The Authorship in the Pauline Corpus,* in *The New Testament in Historical and Contemporary Perspective,* (Eds.) Hugh Anderson & William Barclay, Oxford, Basil Blackwell, 1965, pp. 209-235, tabl.
M. Levison, A.Q. Morton & Dr. W.C. Wake, *On Certain Statistical Features of the Pauline Epistles,* in *The Philosophical Journal,* III, 1966, n. 2, pp. 129-148, tabl.
M. Levison, A.Q. Morton & A.D. Winspear, *The Seventh Letter of Plato,* in *Mind : A Quaterly Review of Psychology and Philosophy,* LXXVII, N.S., 1968, n. 307, pp. 309-325, tabl.

(7) L.A. Sherman, *Some Observations upon the Sentence-Length in English Prose,* in *The University Studies of the University of Nebraska,* I, 1888, n. 1, pp. 119-130.

(8) Bernard O'Donnell, *An Analysis of Prose Style, to Determine Authorship.* The O'Ruddy. A Novel by Stephen Crane and Robert Barr, The Hague-Paris, Mouton, 1970, 108 p., tabl. (= *Studies in General and Comparative Literature,* volume IV).

(9) Louis Tonko Milic, *A Quantitative Approach to the Style of Jonathan Swift,* The Hague-Paris, Mouton, 1967, 317 p., tabl. (= *Studies in English Literature,* volume XXIII).

(10) Erich Kalisch, in *Mitteilungen des Deutschen Germanisten-Verbandes,* XVII, 1970, n. 1, p. 35.

(11) *Calculi,* bimonthly newsletter edited by Stephen V.F. Waite, Department of Classics, Dartmouth College, Hanover, New Hampshire 03755.

(12) *Computers and the Humanities,* editor Joseph Raben, published by Queens College of the City University of New York, 1966.

(13) Yehuda T. Radday, *Isaiah and the Computer : A Preliminary Report,* in *Computers and the Humanities,* V, 1970, n. 2, pp. 65-74.

(14) Philip J. Stone, et alii, *The General Inquirer : A Computer Approach to Content Analysis,* MIT Press, 1966, This work is extensively reviewed by Charles Kadushin, Joseph Lovett and James D. Merriman, in *Computers and the Humanities,* II, 1968, n. 4, pp. 177-202.