NOTES AND DISCUSSION

Dom Jacques Froger

THE ELECTRONIC MACHINE AT THE SERVICE OF HUMANISTIC STUDIES

The multiplication of documentary sources (books or objects) from which the various disciplines glean their data has made the machine an indispensable tool of scholarship today.

The horizons of research are broadening continually. The traditional sciences have widened the field of their investigations. Linguistic studies, for example, are no longer limited to a few languages, but encompass all dialects which have ever been spoken or are spoken today throughout the world. Archaeology is armed with methods of prospecting undreamed of yesterday, and uses the airplane and aerial photography to locate sites. A submarine branch of archaeology, exploring the beds of the seas, has sprung up alongside of land-based archaeology. New sciences have come into being relatively recently, like demography, or

Translated by Nora McKeon.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

even very recently, like experimental psychology. All of these sciences owe something to humanistic studies, not only the applied sciences and industry but even the pure sciences such as mathematics, methodology, heuristics, for scholars are men, integrated into groups, and their intellectual activity falls into the domain of psycho-sociology.

The growth of documentation, moreover, is not a spontaneous development. It is the result of deliberate and almost feverish efforts. Man has become aware of a highly disturbing fact: not only are the data with which the sciences work perishable, but they are actually decaying, little by little, with each passing day. In the domain of linguistics, for example, the dialects of our civilized countries and the languages of "savage" populations are in the process of disappearing. The same is true of all aspects of folklore: customs, music, plastic arts, costumes, etc. The "primitive" civilizations are fading away, and ethnologists find the object of their studies crumbling before their eyes. In sociology observations made one day, in a world as mercurial as ours, are out of date within a few years and enter the domain of history. Archaelogists deplore the fact that many sites have been pillaged by inexperienced hands. While precautions are taken to reserve excavations for specialists, not even the specialists can hope to extract from a site all the information it contains. In archaeology, to discover is always, in a certain measure, to destroy. Many manuscripts were lost through the carelessness of the humanists of the 15th and 16th centuries, discarded as useless once they were published in a printed edition, or split up and scattered at the death of their owners. Even nowadays the care with which old parchments and papyri are conserved cannot protect them from accidents which progressively whittle away the stock. Monuments fall into ruin, the acidity of the air attacks statues. The disasters caused by war are notorious, but even apart from catastrophies, a banal accident, like a fire or a hurricane, would be enough to cause irremediable destruction.

In short, objects of all kinds are inevitably subject to the wear and tear of time. There is thus every reason to collect and catalogue the traces of the past before they disappear. It could even be argued that we are not hurrying fast enough and that centuries to come will accuse us of heedlessness.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

Given the progress of the traditional sciences and the development of the new, the number of documents is augmenting at an overwhelming pace. Each year some two million articles and several hundred thousand books are published, and this output is likely to increase tenfold in a few years. Documentation thus increases in a geometric progression, and proliferates at an exponential rate, that is, its augmentation at each instant is proportional to the state it has achieved, or to an increasing function of that state. It follows a curve which is practically horizontal at first, then rises progressively until it finally assumes an almost vertical direction.[1]

The curve of the increase in documentation corresponds rather well, naturally enough, to those of scientific discoveries and of the number of scholars. As Robert Oppenheimer remarked (or at least is said to have remarked), "9/10ths of the scholars which humanity has fathered since its birth are living today. We have made more progress in 40 years than in 40 centuries. 99 per cent of our knowledge is due to men who are still alive."[2] The world population, as well, is increasing more rapidly, so that "documentary demography," "the demography of discoveries," "the demography of scholars" and plain and simple "demography" march on at about the same pace.

In documentation, as in other domains, it would be unreason-

[1] Extremely interesting information concerning the increase in documentation can be found in R. Caude and A. Moles (with several other collaborators), *Méthodologie vers une science de l'action* (Paris, Gauthiers-Villars, 1964), particularly, in the chapter "Sociologie de l'action" by H. Migeon and A. Moles, the graphs of the increase of scholars in the major parts of the world (p. 192) and of the rhythm of important inventions since the 10th century (p. 195), and in the chapter "Mise en ordre des connaissances" by J. Dubas, which has a paragraph on the "Développement de la production documentaire" illustrated with a graph (p. 276). In fields such as those which we are treating here, the term "exponential" should not be taken strictly. Explaining the characteristics of a geometric progression (p. 35) A. Moles notes, "when the curve is relatively lacking in precision, it is often difficult to judge whether it is a parabole $y = Ax^2$ or an exponential $y = e^{kx}$."

[2] This is true only in the domain of science in the proper sense of the word, where progress depends solely on observation and calculation. The case is completely different in purely qualitative fields, such as the arts. Here it is obvious that the number of masterpieces and of great artists which have come into the world in the past 40 years is miniscule in comparison with all that the past has produced.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

able to consider this exponential growth alarming. With curves of this type it is dangerous to extrapolate. Unforeseeable events can intervene and modify the pace of evolution.[3] From the 7th-6th century B.C. to the onset of the Christian epoch, Greek science seems to have followed an exponential curve like that which we are experiencing today. Then the arrival of the Roman "barbarians" lowered the cultural standard and slowed down progress. Finally the other "barbarians," profiting from the weakness of the Romans, invaded the Empire and caused a sort of collapse. Taking off again almost from zero in Western Europe, the curve of important scientific discoveries becomes regular only in the 10th century and reaches the level which the ancient Greeks had attained towards the 17th-18th century. Its climb, practically vertical today, might bend towards the horizontal if some type of saturation occurs. An atomic world war would knock it down, as did the succession of barbarian invasions in the first millenium of the Christian era. But even if no catastrophy intervenes, the present vertiginous increase in the number of books, sciences, discoveries, scholars, and uncovered vestiges of the past might not conserve its exponential pace indefinitely.

*

Whatever the distant future may hold, the prodigious expansion which we are witnessing today, and undoubtedly will witness for some time to come, has pushed the problem of documentation to a position of first importance. The already large and still growing quantity of material that the various sciences bring into

[3] In the work cited above (note 1), H. Migeon and A. Moles illustrate (p. 193) "the dangers of forecasting" by presenting the curve of the number of workers in the electrical construction industry in England, with this reflection: "It is obvious here that it would be dangerous to go too far in extrapolating an exponential curve; the absurd result would be that in 1990 the entire British labor force would be engaged in electrical construction, while it is now evident that the influence of automation will brake this curve by bringing about an employment saturation." It is highly significant that the curves which the same authors offer to represent the growth of the scholarly population in the world (p. 192) have all begun to lean towards the horizontal (except that of China, which is too recent): "the tendency towards saturation is already becoming evident."

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

play must first of all be organized, according to an at least approximative classification.[4]

Books are collected in libraries. If there are relatively few of them, they can be arranged on the shelves according to subject matter. Yet this procedure, although highly convenient, is not entirely satisfactory, since many books deal with several disciplines at the same time. Moreover, it is impracticable in large libraries, where circumstances determine the accumulation of stocks, in such a way that the organization of the books is largely arbitrary. In any event, a library cannot be used as is: it must be doubled by a "phantom" library or catalogue.

The component of the catalogue is not the book but a descriptive account which gives the characteristics of the book: author, title, publisher, date, etc. In practice, unfortunately, a catalogue never represents the entire contents of a library. It leaves out magazine articles and monographs published together as part of larger works. It notes the existence of works published in the form of collections but does not enumerate their contents, for the analysis would be too time-consuming a process.

The catalogue can take the form of a list or of a card-index. The list, a bound volume, is compact and easy to publish, but has the disadvantange of being a "closed" catalogue, which stops with the date of its edition. To bring it up to date it must be completed by a supplement, which must later be supplemented itself, and so on *ad infinitum.* The card catalogue, on the other hand, is "open". It can be brought up to date continually simply by inserting new cards in the proper place.

Whatever its form, a catalogue presents another problem: in what order will the elements which compose it be arranged? The simplest method is that of alphabetical order by author (or by title in the case of anonymous works), but this solution creates a difficult situation for the researcher who wants to make a bibliography. He is presumed to know in advance which authors have dealt with the problem which interests him, in which case he would hardly be hunting for material to begin with. The only way to avoid this vicious circle would be to provide a systematic

[4] Cf. J.-C. Gardin, "Problems of Documentation," in *Diogenes*, No. 11, Fall 1955, pp. 107-124.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

classification by subject matter, but this is such a painstaking task that it is often left in a rudimentary state, even in large libraries.

Transportable objects are grouped in museums or collections. They are easier to arrange "on the shelves" according to their nature, origin, date, etc., but the problem of a catalogue arises here as in the case of books, although under different conditions. The descriptive card lists the objects' principal characteristics according to a preestablished "code." Objects which cannot be transported, like monuments, are obviously left where they are, and only their "phantoms" can be assembled in repertories.

Since the number of objects which the past has bequeathed us is finite, it would be possible to draw up a complete list of them and to establish a corpus which would assemble in one spot all that is scattered throughout the world. For example, the number of ancient and medieval manuscripts is finite and, if they have not all been examined yet, we are at least certain that no new ones will come to light. A complete enumeration could therefore be made. Mr. Lowe, in the United States, has published in his *Codices Latini Antiquiores* a corpus of all Latin manuscripts up to the 9th century. In France, Mr. Samaran and Mr. Marichal establish a corpus of dated manuscripts posterior to the 9th century in their *Catalogue des manuscrits portant une indication de copiste, de lieu ou de date.* A repertory of all Latin manuscripts without exception will undoubtedly be drawn up one day. The Greek manuscripts, far fewer than their Latin counterparts, are much easier to catalogue comprehensively, a task currently undertaken by Father Richard at the Institut de Recherche et d'Histoire des Textes (CNRS, Paris, 15 Quai Anatole France). The same sort of enumeration is applied to all types of objects. The Swiss paleologists Mr. Bruckner and Mr. Marichal have established, in the *Chartae Latinae Antiquiores*, a corpus of all Latin papyri anterior to the 9th century; Mme Gauthier, at Limoges, is drawing up a corpus of meridional enamels; the *Index of Christian Art* prepared by the University of Princeton in the United States comprehends everything concerning Christian art and iconography; a corpus of incunabula has been established in Germany by Hain. Since the number of publications which have appeared since the 16th century is finite at every moment, it

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

would be possible to conceive of a "universal" card-index which would cover the contents of all the libraries throughout the world and, constantly kept up to date, would report at every instant on the state of man's book learning. The constitution of corpora is one of the characteristics of our time. These collections, of course, will never be absolutely exhaustive, but an effort is made, at least, to make them as complete as possible.

Once the information has been gathered, the next step is to use it. The first stage in a study is drawing up a bibliography on the subject, to find out how far research has already been carried and what points remain to be tackled. The problem of establishing a bibliography is not the same in all disciplines. Those of a " scientific " nature, whose goal is to solve new problems by formulating laws, are interested only in recent works, in the latest discoveries and the books which describe them. A publication loses all its interest as soon as another one restates the question. In these fields bibliographies shed out-dated material as the science progresses. They renew themselves by a process of constant substitution and vary in composition rather than quantity. The situation is different for the "historical" disciplines, where an effort is made to reconstitute the past by tracing the evolution and the sequence of events. The bibliography cannot neglect anything. Everything which has been written has documentary value, and no book can be thrown out. Yet, though the difficulties of documentation are not always the same, they are equally imposing in all disciplines.

The classification of documents in libraries, museums and corpora is only approximate. Thus manuscripts are arranged in corpora by order of the libraries in which they are to be found. In order to have an index by dates, sources, types of handwriting, etc., the pages must be cut out and the publication transformed into a card-index. Mr. Samaran and Mr. Marichal have provided for this possibility by publishing the "dated manuscripts" in the form of unbound sheets, printed on one side only. Reviews such as *Scriptorium* or the *Revue d'Histoire Ecclésiastique* in Belgium offer researchers "bibliographies," but they too can be used conveniently only by cutting them up into card-indexes. Why should we continue to use the medieval system of lists today? Would it not be better to decide once and for all to publish bibliographies

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

in the form of index cards, possibly in several copies, so that each researcher can arrange them as he preases?

In short, the documentation at our disposal today is extremely plentiful, poorly organized, and augmenting at a stupendous rate. We are crushed under a mass of material, so prolific that it can no longer be processed "by hand": we must call in the aid of machines.

*

But which machines? There are three categories: the traditional card catalogue, the mechanical machine, the electronic machine.[5]

The traditional card-index, with its cards arranged according to a hierarchical system of key-words or subject matter words allowing for numerous classifications, is already a machine; it is simply constructed and manipulated by hand.

The mechanical machine is basically a sorter which handles the cards. It can be powered by any source of energy, and could be made to operate by turning a crank. In practice it is run by an electric motor and is thus an "electric machine," or one in which electricity provides only the propulsion, while the operations are carried out "mechanically" (by grooves, gears, levers, etc.). To enable the machine to take hold of and handle the cards, any information written out on the cards is represented, according to certain conventions, by perforations within the cards, notches on the sides, etc.

The electronic machine differs from the electric machine in that electricity no longer plays simply the extrinsic role of propelling a mechanism which executes the work, but actually carries out the operations itself. The operating agent is the flux of electrons which makes up the electric current; the instrument of work is the magnetic impulsion produced by the electrons; and the "mechanism," in such a machine, plays only a secondary and

[5] The reflections which follow, and a good number of those included in this article, draw particularly on a lecture given by Mr. J.-C. Gardin on March 18, 1965 at the center of the Société d'Encouragement pour l'Industrie Nationale (44, rue de Rennes, Paris) and on a conversation which he was good enough to grant me a few days later. I thank him for his kindness and hope that I have not distorted his ideas here.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

auxiliary role. The information (letters or numbers, conventional signs, etc...) enters the machine in the form of perforations, according to a preestablished code, either on cards or on bands of paper. These perforated papers slide between a conducting surface and a system of "brushes" made of steel blades. The non-perforated portions interrupt the electric current, while the perforations let it pass through. The language of the machine is thus binary and works on a basis of "all or nothing," signal or absence of signal. This is why the machine's calculations are carried out either in a purely binary system (numeration based on 2) or in a decimal system coded binarily (numeration based on 10 in which each number is coded in binary form). The information is recorded on tapes and introduced into the "memory" of the machine, which can take different forms: drums, disks, etc. The operations which the machine should carry out are introduced in the same perforated form as the information to be handled, and the two together constitute a "program". At the end of the operation the results are printed in ordinary letters and figures, or even in the form of graphs or diagrams. Alongside of the "digital" machines (from the word "digit"), which operate on figures or discontinuous quantities, there are "analogical" machines which operate on continuous quantities, somewhat like slide-rules. Progress is rapid in this domain of technology. A computer is obsolete within a few years, and the future should hold increasingly amazing achievements.[6]

[6] The electronic machine is the fruit of a sort of collaboration among England, the United States and France. Its true distant ancestor is not the calculating machine devised by Schickard (1624), Pascal (1645) or Leibniz (around 1700), which is more in the line of office machines, but rather the Analytical Engine designed by the Englishman Charles Babbage around 1830. Conceived ahead of its time, this apparatus was never constructed because the idea was too advanced for the technical means which industry then had at its disposal. Babbage's idea was revived in 1937 in the United States by Howard H. Aiken, who had the International Business Machine Corporation (IBM) construct an electronic machine called the "Mark I." The idea of a machine powered by electronic resources was expounded in 1938 by the French professor Louis Couffignal. Aiken's second machine, the "Mark II," was electromagnetic and was out dated even before it was produced by the E.N.I.A.C. (Electronic Numerical Integrator and Computer), the first truely electronic machine, built by J.P. Eckert and J.W. Mauchly at the University of Pennsylvania around 1944. England followed close on the heels of

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

Electronic machines today come in all sizes, from the small tabulator to highly powerful computors. It would be a mistake to think that the most complicated version is always the most profitable in all circumstances. In choosing a type of machine for a particular job two points must be considered: speed and price, or outlay in time and in money. The solution to be adopted is the one which seems most economical in all respects.

Cards can be handled more rapidly by a mechanical machine than by hand. The electronic machine works even faster, and it is perhaps this element of speed which strikes us as most incredible. It is an exaggeration, of course, to say that it works "at the speed of light." The manual preparation is sometimes quite lengthy, even more time-consuming than the calculations themselves. There is a certain amount of lost time and some operations which are carried out mechanically: the unrolling of the tapes, rotation of the memory drums or disks, printing of the results, etc. Despite these limitations, the machine executes calculations at a speed which is out of all proportion with manual work. For example, the IBM 7090 computor (built in 1961) can carry out 229,000 additions or subtractions in one second. To calculate the difference between the theoretical values and the measured values of the field of terrestrial gravitation would require: by hand, with pencil and paper, 1000 years; with an office calculating machine: 5,000 weeks; with an electromagnetic computor

the United States; a project was under examination in 1946 and the machine constructed in 1951. In France, the Compagnie Bull built machines based on the ideas of Professor Couffignal. On the history of electronic machines, particularly in the United States, see: El. Berkeley, *Giant Brains, or Machines that Think* On the principle and operations of the electronic machine, see, for example: N. Chapin, *An Introduction to Automatic Computers* (New York, Van Nostrand, 1953), and in the "Que sais-je?" collection (Paris, Presses Universitaires de France): J. and J. Poyen, *Le langage électronique* (1960); P. Demarne and M. Rouquerol, *Les ordinateurs électroniques* (1959); B. Renard, *Le calcul électronique* (1960). The best description of the various forms of the modern electronic machine is the work of François Gauchet, Roger Lambert and Jacques Violet, *Le calcul automatique en psychologie* (Paris, Presses Universitaires de France, 1965, in the Collection "Le Psychologue," No. 22). The text is illustrated by some thirty diagrams, 15 photographs and 10 tables; the last part offers a concrete example of the use of automatic calculations in psychology. The authors have succeeded in giving explanations which are both extremely detailed and yet understandable to the non-specialist.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

(1944): 3,750 days; with an electronic IBM computor (1948): 50 hours; with an IBM 704 computor (1957): 75 minutes.

On the other hand, work executed by the electronic machine is expensive. The cost range is different in the various phases of operation. The expenses involved in perforating cards or bands at the beginning, and printing the results at the end are about the same for all machines. The cost of studying and preparing the programs increases in proportion to the strength of the machine. The expenses involved in the treatment itself are inversely proportional to the size of the machine, that is the more powerful the machine, the less it costs to perform an elementary operation. The output of the large machine is thus superior to that of the small. The price of a machine is obviously higher the larger the machine, and the same is true of rental fees; let us say, to give an idea of the price scale, that one machine-hour would cost: on a small IBM tabulator: $4; on a more powerful machine: $20; on a Gamma 50 Bull: $100; on the largest machines that exist today: $1,000. But, since the rise in the price of a machine-hour is compensated by a saving of time and a superior output, it is simply a question of choosing the solution which best balances these different factors. Lengthy calculations are much cheaper on a very large machine than on a small one; on the other hand, it is not advisable to make a machine work too far above its maximum power. In choosing a machine, the problem is thus to find the model which is perfectly proportioned to the job at hand.

It must be added that the electronic machine is often not only more expensive, but even less rapid than the mechanical machine[7]

[7] In the field of philology, a good number of projects remarkable for their breadth and precision are carried out with the aid of the mechanical machine as well as the electronic computor. This is the case particularly at the Laboratoire d'Analyse statistique des Langues anciennes (University of Liège), directed by Mr. Delatte and Mr. Evrard, who describe their work in an article in *Antiquité Classique*, vol. XXX (1961), fasc. 2, pp. 427-442. Procedures where mechanical machines play an important role have recently been adopted by Mr. Paul Tombeur to study the language and style of the medieval chronicler Raoul de Saint-Trond. He explains his methods in an article entitled, "Application des méthodes mécanographiques à un auteur medieval" (in *Archivum Latinitatis Medii Aevi*, vol. 34, 1964, pp. 125-160. His work will be published in the near future under the patronage of the Commission Royale d'Histoire de la Belgique.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

which, in its turn, can very well offer no advantage over manual work. Experiments conducted with this point in mind have shown that the processing of a number of cards up to several thousand can be carried out more advantageously by hand, on a table top, comparing the cards by transparency or some equally simple method. The job can be finished in a few hours and with a minimum of expense in this way, and would require far more time and money if it were handled by a machine.

In short, it is simply common sense to do a bit of "operational research" before executing a task and to choose the process best suited to the circumstances, without being ashamed of manual work when it proves to be the simplest solution.

*

Let us suppose that the massive quantity of information justifies the use of an electronic machine, on the understanding that our observations will naturally apply, with the necessary modifications, to mechanical machines as well. The differences between the card-index that is then established, in "automatic documentation,"[8] and the traditional variety set up by hand are more quantitative than qualitative. The former contains more information and thus offers more extensive possibilities of combination and classification covering a larger number of elements. But, in the last analysis, it too is made up of key-words arranged in a hierarchical fashion.[9]

In automatic documentation, different methods must be used for "literal" and "non-literal" material. The latter are objects; once their characteristics have been transcribed into code the processing is much like calculations. In the case of "textual"

[8] Concerning automatic documentation we will cite only J.-C Gardin, "Etat et tendances actuels de la documentation automatique," in the review *La Traduction automatique* (Paris, 5th year, No. 1, March, 1964). In his notes the author gives a considerable bibliography concerning work achieved in the United States and in France.

[9] The continuity of what is known as problems of "classification," from traditional documentation to modern documentation, is set forth by B.C. Vickerey, *Classification and Indexing in Science* (London, 1959), and by de Grolier, *Etude sur les catégories générales applicables aux classifications et codifications* (Paris, 1962).

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

information, or texts written in a "natural language" (a language as the linguists understand it: French, German, Arabic, etc.), the documentary research, or collecting of documents, is followed by a more delicate operation: the "documentary analysis," which condenses the text written in a natural language into a "documentary language" or "language of information." This is achieved through a process of indexing which collects and combines two elements: on the one hand a list of key-words which, when complete, constitutes a "glossary of documentation" including all the technical words employed in a given discipline; on the other hand, a series of symbols which represent the semantic and syntactical relationships of the key-words. It is obvious that a résumé or a list of key-words cannot condense the entire contents of a book or article and retains only the essential. A certain amount of information is thus lost at this stage, but this is a drawback which must be accepted as inevitable. Once this indexing in a documentary language has been accomplished, the machine records the results in its own symbolic form and its own language.

The lists of key-words resuming the contents of books or articles make it possible to draw up bibliographies by classifying under one heading all the works which treat a given subject. The difficult part of the operation, however, is to draw up such lists. Various methods of rapid automatic indexing have been tried. But the title does not suffice, since it is often too vague and does not really identify the subject. Nor can the index be based on the words used most frequently; experience has shown that they generally bear no relation to the subject of a work. Another method, by citations, has been tried: assuming that an author normally cites the books or articles which have treated the same topic before him, a list of works which cite each other would constitute a bibliographical ensemble concerning a given subject. Such a method does not seem likely to produce satisfactory results. Generally speaking, all of these rapid processes, which attempt to "short-circuit" the operation of indexing, evade rather than solve the problem. None of them gives complete satisfaction. After various unfruitful tentatives, researchers today must still content themselves with résumés and indexes drawn up by hand by an intelligent reader. Completely automatic documentation will

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

perhaps be achieved one day, but it cannot be expected in the near future.

*

The problem of automatic translation is posed already the minute that key-words are fed into the machine. Moreover, even the most perfect machine-produced documentation will not satisfy the needs of researchers until it has been crowned by automatic translation. Once the bibliography of a subject has been drawn up, it remains to consult the works it mentions. But those which are written in a language with which the reader is not familiar are as inaccessible as if they did not exist. He is obliged to wait until they have been translated; but competent translators are hard to find, they work slowly and their services are expensive.

Only a translating machine could overcome these highly annoying barriers and ensure that all important works be placed without delay at the disposition of all researchers, whatever their nationality. Under such conditions the various disciplines would advance more rapidly and more easily.[10]

The first scheme for automatic translating was devised in 1933 by the Russian Smirnov-Trojanskij, but neither his country nor any of the others picked it up. In 1946 the Englishman A.D. Booth, who was probably unacquainted with the work of his Russian predecessor, worked out the idea of automatic translating in his turn. He proposed to Warren Weaver, of the Rockefeller

[10] Here is an example, not very recent but significant nonetheless, of the obstacles to the progress of science which the scarcity of translators creates. In the *Histoire générale des Sciences* published under the direction of R. Taton (Presses Universitaires de France) vol. 2, in the chapter concerning the "Naissance de la chimie moderne," Mr. Daumas adds the following footnote (p. 553) to his discussion of Lavoisier's work around 1770: "Much before that period the Russian scientist M.A. Lomonossov published work which included remarkable anticipations of later discoveries, among them the conceptions of the atomic constitution of bodies and of kinetic energy due to molecular agitation... Unfortunately his writings, published in Russian, never came to the attention of chemists from other countries. There is no mention of them in the English, French and German literature of the time. A broad diffusion of his ideas would undoubtedly have promoted the future of modern chemistry."

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

Foundation, to build a translating machine and his suggestion was accepted enthusiastically. At first Booth thought of automatic translation as a process analagous to deciphering a coded message, but he soon realized that the problem was far more complicated and pertained above all to the domain of linguistics. At that time this discipline had made great progress, from which Booth's and Weaver's research benefited. Electronic scientists and linguists worked together closely and in 1952 held their first conference, in the United States, to study together the problems of translation by machine. In 1954 the first experiment in automatic translation, from Russian to English, was held at Georgetown University. These two languages received the highest priority since they could reasonably be considered as the most representative of the "East" and of the "West" and since the first and most assiduous work was done by English-speaking researchers. Not before 1955, the year in which William Locke and A.D. Booth published the first book on the question, did the Russians enter the field to catch up with the Americans. Since then, research has been initiated not only in the United States, England and Russia, the three leading countries in this domain, but also in France, Italy, Scandinavia, Japan and, in short, in all countries of an elevated cultural level.[11]

[11] In the United States there are at least ten centers dedicated to automatic translation, notably at Harvard, Georgetown, Berkeley, Los Angeles, and an Association for Machine Translation and Computional Linguistics. In England, Birckbeck College (Department of Numerical Automation) and Cambridge (Cambridge Language Research Unit) can be cited. In Russia the most important body seems to be the Experimental Laboratory of Automatic Translation at the University of Leningrad. In France there is a "Centre d'Etudes de la Traduction Automatique" at Paris and another at Grenoble; the Association pour l'Etude et le développement de la Traduction Automatique (A.T.A.L.A.), 20, rue de la Baume in Paris, publishes a review, *La Traduction Automatique*, which is international in its coverage.

Concerning the history and the technique of translation by machine, the best overall discussions are: Emile Delavenay, *La Machine à traduire*, (Collection "Que sais-je?," Presses Universitaires de France, 1959) and A.G. Œttinger, *Automatic Language Translation: Lexical and Technical Aspects* (Harvard University Press, Cambridge, Mass., 1960). For details on recent research, above all in the United States, but also in England, France, Italy and Japan, consult the communications presented at the First International Conference on *Automatic Language Translation and Linguistic Analysis*, held at the National Physical Laboratory, Teddington (Middlesex), from September 5 to 8, 1961. The reports and discussions have been

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

Despite considerable efforts throughout the world, automatic translation has not yet become a reality. As a result the general public feels a sense of deception when it realizes that the translating machine, which it thought was already or would soon become a working reality, is not yet in operation and most likely will not be in the near future. The specialists, on the other hand, although they also went through a stage of somewhat exaggerated optimism, have always been more reserved in their prognostics. They have become even more so today, since their task appears increasingly difficult as work progresses. Their prudent attitude, however, is not an omen of failure. It is true that the experiments reported in reviews and collections of monographs leave a first impression of confusion, but this is simply an outward appareance. The specialists are divided on a number of points, but they do agree on the general direction which must be taken. Automatic translation is not really at an impasse, as is sometimes alleged; at the most it is in a state of "stagnation." Moreover, the slow rate of progress is hardly surprising considering the difficulties which must be overcome.

When a text is to be translated by a machine it is first fed in, as usual, in the form of perforations. At the same time the machine is given dictionaries and grammar rules. It must recognize, in the original language, the meaning of words and their syntactical relations, then find, in the second language, the words which have the same meaning and arrange them according to the proper syntax. Of these two stages the second, it seems, is not the more delicate: once the machine has assimilated the vocabulary of a language and the rules of its syntax it can construct correct sentences without too much difficulty. The operation which the machine finds trickiest is to "understand" the original language, for the simple reason that we have not yet succeeded in formulating discursively, for the machine's use, the mental

published in English by Her Majesty's Stationary Office (London, 1962). The fourteen most important communications, among the thirty-six presented at this international conference, have been translated into French and published under the title: *Traduction automatique et Linguistique appliquée* (Paris, Presses Universitaires de France, 1964). This volume is the first in a collection which will appear under the auspices of the A.T.A.L.A.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

operations which we often carry out intuitively when we read. The grammatical function of words in a sentence is not always evident in the light of precise rules, and we deduce it from the general context. In languages without declensions, like French, English, Spanish, etc., the subject and complement are generally distinguished by their positions in relation to the verb; there may be inversions, but they do not bother us since we are guided by the meaning of the sentence. If, for example, we find the words "gnaw," "bone," "dog" we know, whatever the order of the words, that it is the dog that gnaws the bone and not the bone that gnaws the dog. But how to explain this to the machine? Semantic difficulties are perhaps even more serious. They are due above all to polysemous words, or words which have more than one meaning, the most common, unfortunately, being those which have the greatest variety of meanings. Homonyms, or different words which are written in the same way, help to confuse the machine, as do the "lexies," or groups of words which form a whole (e.g. *with reference to*) but can be split up by modifiers (*with particular reference to*, or *with reference to...and to*).

Each language presents its own special problems. German, for example, has compound words which must be "decomposed" by the machine, a task which is particularly delicate when a word changes meaning according to whether, in the process of cutting, the letter which makes the difference is attached to the first or the second half of the compound word. Thus *Wachtraum* means "waking dream" if it is cut *Wach/traum*, and "guard room" if it is cut *Wacht/raum.* In Spanish a way must be found to distinguish the pronoun complement from the verb to which it is joined: for example *dale*, "give him," which the machine must interpret *da/le.* In order to resolve semantic ambiguities, the machine is asked to examine the context; thus the French word *pièce* would be translated by *gun* in English and *cañon* in Spanish if artillery is being discussed, by *coin* in English and *moneda* in Spanish if the subject is finance, by *room* in English and *habitación* in Spanish if the context speaks of a house, etc. Nonetheless, the difficulties are such that the translating machine still makes mistakes about two thirds

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

of the time. Far from being exceptions, as optimistic authors claim, ambiguities are thus frequent occurrences.

Faced with difficulties of such magnitude and anxious to get results more quickly, researchers had searched for ways of sparing the machine part of its work by doing it manually. These methods involved either solving the ambiguities in the original language in advance through a "pre-edition" or preliminary indexing, or solving them in the final language through a "post-edition" which chose among the various solutions proposed by the machine. Such procedures have been abandoned today. As Mr. Gardin says, attempted short cuts in automatic translation, as in documentation, are a waste of time and divert attention from the real problem. It is far better to stick to the main road and attack the difficulties head on in order to find their theoretically correct solution. There is no reason to be pessimistic or discouraged. "If the human spirit can perform certain operations, there is no reason why a machine cannot be made to perform them." All that must be done is to analyse carefully the steps which the human mind follows in reading and understanding a text, and to push linguistic studies still further. This will take time. Automatic translation, according to current predictions, will not be realized for about fifteen years. Let us be patient, then: "We have waited for centuries," Mr. Gardin remarked with a smile, "We can well wait fifteen or twenty years more."

It should be noted that the constructors of translating machines have moderated their ambitions. They have given up the idea of making machines able to translate anything and now think simply in terms of machines specialized in a certain discipline, with an appropriate dictionary. Hopefully, this restriction is only provisional, since some books are works of synthesis and many of the humanistic studies are "inter-disciplinary" by nature: sociology or demography, for example, draw on history, law, psychology, medicine, mathematics (statistics), etc. at the same time. In any event, it goes without saying that automatic translation will never replace "literary" translation, since it could never convey fine points of style; all that is asked of it is to translate the meaning exactly.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

*

The most highly vaunted advantages of the machine are speed and security. We have already spoken about the first; let us now turn to the second. The electronic machine is said to be infallible. This is absolutely true, but only if certain precautions are taken.

Errors which can be laid to the machine are no longer to be feared: the methods used to avoid them are so efficacious that machine-made mistakes have been practically eliminated. When errors do occur they are most often due not to the machine but to the man who uses it.

Man can make errors of logic while posing the problems, working out methods or interpreting results. The machine is obviously not the guilty party. It carries out a job, but does not evaluate the method imposed on it, it answers questions, but does not judge whether or not the questions are pertinent; it delivers facts, but draws no conclusions. We will come back to this subject when we speak about the services which the machine can render in various domains, for it can be turned to best account only through judicious use.

Material errors can occur each time that man must intervene. To reduce the chances for error, automatization is applied to as many operations as possible so that, once set in motion, they roll along of their own accord. But human intervention cannot be dispensed with in the preparation of material and the initiation of the operations, and it is here that errors must be watched for. Actually, the delicate point is not the program, since the machine checks everything while executing it and points out incoherencies or oversights if it finds any. The danger of mistakes lies principally with the information, and the perilous stage is that of perforation.

The cards or bands are perforated on a large electric typewriter. This machine is set up and works like normal typewriters, but the keys are connected to another apparatus which punchs the perforations according to a predetermined code. Since the secretary reads the text in typescript as she perforates, she does not work "in the dark"; however, like all typists and, more generally, all copyists, she can and does make mistakes. The

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

work must therefore be checked. This is done by a specially conceived machine called the "verifier" which functions in the following manner: the band or series of cards is placed in the machine, a second perforation is made over the first and, if they do not coincide, the verifier refuses to continue. The point at which it stops is where a mistake was made during one of the two perforations; the error is checked and, if it occurred during the first perforation, is corrected. The process of correction is easier if cards are used: the faulty card is remade, and an insert card is added if necessary, bearing the same number as the preceding one, but with a special mark to distinguish it. It is generally held that the possibilities of error during the first perforation are less than 1 per cent; they are the same during the second. Once a text has passed through the "verifier," the possibility of error should thus have been reduced to about one in 10,000, a negligible percentage.

Experience shows, however, that these estimates are perhaps overly optimistic. In 1960 we tried to use an electronic machine at the Compagnie Bull for the operations of textual criticism which philologists do by hand. The first step was to collate several manuscripts of one work in order to isolate, by an "automatic" comparison, their differences in tenor or "variants."[12]

The second perforation by the "verifier" did not lower the percentage of errors in the slightest; on the contrary, it raised it, since there were 108 mistakes (erroneous signs) after the checking process and only about 74 after the first perforation. About thirty mistakes made during the first perforation had been repeated during the second (a curious fact, of interest to philologists), and nine of them constituted important blocs, since they consisted in the omission of two entire words (*fecit* and *casu*). One correct card had been repeated by error, for a total of 48 erroneous signs in one blow. When extra cards were inserted words were cut improperly (faulty spacing) practically every

[12] A detailed description of this experiment has been published in the *Bulletin d'information de l'Institut de Recherche de d'Histoire des Textes* directed by Mr. Glénisson (Paris, 15, quai Anatole France; CNRS) vol. 13 (1965); Dom Jacques Froger, "La collation des manuscrits à la machine électronique." The program of comparison is the work of Mme Renaud.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

time that the occasion arose. Finally, a number of cards that were redone to correct one error introduced another in a different spot, as often happens in printing houses when linotype is used. The redone cards represented a first run which should have been checked by passing the whole text through the "verifier" once more.

Though the percentage of mistakes left by the machine was low, there were enough of them to throw the results off entirely. The idea of a "tolerable margin of error" is perhaps admissible in circumstances where it is certain that the mistakes are trifling and cannot prejudice the work; but, in principle, the information fed into the machine must be absolutely accurate. Among the advantages of automatic treatment, security outweighs speed; speed would even be valueless if the results were not completely correct.

In order to eliminate the errors left by the machine, I had to turn to the only truly reliable method, which is standard procedure particularly at the Section d'Automatique Documentaire de Marseille directed by Mr. Gardin: to compare, by hand, the texts produced by the machine with the originals and continue to ask for new proofs, as in a printing-house, until everything was perfectly in order. The corrections which I demanded were, in fact, carried out impeccably on the first try. I wonder, therefore, if it would not have been simpler to give me the results of the first run: there would have been fewer mistakes to correct. Might the "verifying" machine, with the false sense of security it gives, be more harmful than helpful? Only the specialists can judge; they know, at any rate, that the verification is reliable only if carried out at least twice.

✱

One of the advantages of the machine, and not the least important, is not generally emphasized enough: its total lack of intelligence. Incapable of working out a method, it does what it is told, nothing more, nothing less, in a purely material fashion. Unable to think, it obliges the man who uses it to make a greater and more careful mental effort in analysing the operations which the machine will be asked to perform. When work-

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

ing by hand the scholar should, in principle, always stick to a rigorous methodology; but in reality he sometimes circumvents difficulties and contents himself with an approximation. The machine inexorably forbids evasion and refuses to tolerate imprecision. It requires an exact definition of the givens of a problem and a forecast of all the eventualities, all the difficulties, however trifling, which might arise. This is a precious quality; even if this were its only advantage, the machine would still do its users a great service.

In the field of archaeology, for example, Mr. Gardin wanted to use the machine to classify the multitude of objects found in excavations. He was first obliged to draw up a highly detailed catalogue of all the distinguishing characteristics they presented.[13] Once this was done, he could tackle the classification itself, a task less simple than it may seem at first, since all the characteristics on the list must be sifted in order to find those which should be given the greatest weight. Mr. Gardin will certainly succeed; his work has already left the stage of preliminary study and entered that of experimentation. But even if the difficulties of the classification turned out to be insurmountable and he were forced to abandon that part of his project, the fact alone of having been obliged to draw up a reasoned inventory of the museum pieces would already have represented a highly appreciable achievement. Even if his work stopped there, an attempt to use the machine would already have proved profitable.

The machine offers the same advantages in many other fields. Thus the Corpus of Christian Art undertaken at Princeton University has now reached such dimensions that it has become necessary to process information by machine. But at this point it has become evident that the catalogue drawn up by hand is not sufficiently precise. It will be necessary to redo it entirely, and the directors of the project intend to do so, defining far more objectively, in documentary language, the elements which will be classified by the machine. To get a clear idea of tne ambiguities that can come up in iconography let us take a very simple example: a picture or a sculpture represents a

[13] Mr. Gardin describes his work in an article, "Cartes perforées et ordinateurs au service de l'archéologie." (Review *La Nature*, Nov., 1962, pp. 449-457).

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

woman holding a severed head in her hands; some see it as Judith and Holopherne, others as Herodias and John the Baptist. Which interpretation should be favored? Or again, how can one decide whether a certain church is an example of Gothic or of Romanesque style? A man who is drawing up a catalogue by hand solves this type of ambiguity through intuition, good sense, experience, but all of this is insufficient. Objective criteria must be found, for the machine will not run on approximative information.

We tried in 1960, at the Compagnie des Machine Bull, to use the electronic machine on a philological task which consists in classifying manuscripts according to their variants (tracked down by the automatic collation described above), in order to trace their geneological relationships.[14] This automatic textual criticism will, it seems, render great services to the philologist by sparing him all the work of calculations. But even if this method had never moved from the laboratory to practical application, I would not have wasted my time in trying to work it out, since the use of the machine obliged me to clarify a number of ideas which had been vague before then, and to formulate in logical terms methodological points which had seemed to belong to the domain of pure intuition.

*

Being a computor the machine comes into its own whenever figures and statistics must be handled. This amounts to saying

[14] This experiment was described at the Colloque International de Lexicographie held at Besançon in June, 1961, with a demonstration on a Bull machine at the Faculty of Letters of Strasbourg. Concerning the method followed "by hand" see: Dom Jacques Froger, "La critique textuelle et la méthode des groupes fautifs" (Report presented at Besançon), in the *Cahiers de Lexicologie*, No. 3, 1962, published by the Faculty of Letters and Humanistic Studies of Besançon (Dr. M. Quémada). The program designed for the machine is the work of Mr. Philippe Poré. It was explained by Mme Poyen and Mr. Poré at the second Congress of the A.F.C.A.L.T.I., in October, 1961. For an overall idea of the question, the following work could be consulted: Dom Jacques Froger and Philipoe Poré, *La critique des textes et son automatisation* (Paris, Dunod, to appear in 1966, in the Collection "Initiation").

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

that it is helpful in all domains, since there are hardly any fields today which do not involve mathematics. Here the machine offers the advantages which we have noted: rapidity, security, necessity to spell out information and methods carefully and precisely.

Let us take demography for example, and the apparently simple operation of census-taking. A census involves above all a count of the total number of inhabitants of a country. But in the process, information which will serve as material for statistics is also recorded: sex, age, profession, family situation (married or single), etc. The operation is extremely cumbersome, since it involves dozens or hundreds of millions of people (about 200 in the United States, 55 in Great Britain, 50 in France, 32 in Spain, etc...) If mechanical means are used it takes several years to work out the complete results of a census. In such a case the electronic machine obviously offers the only way of getting the data to sociologists before it becomes at least partially obsolete.

The situation is the same in experimental psychology and in psycho-sociology, where the examination of information collected through questionnaires is a very heavy task. The large number of electronic machines which the American sociologists have at their disposal undoubtedly accounts for the impressive work they have done and for their lead in the field of public opinion polls.[15] Following their lead, the psycho-sociological disciplines in other countries are making wide use of computors, as does, to cite just one example, the Laboratoire de Psychologie Sociale at the Faculty of Letters and Humanistic Studies of Paris.[16]

In these fields the methodological precautions which we described above must be observed with particular care. Figures are used, and they should be, for no work can be scientific

[15] Concerning the methods and results of censuses and public opinion polls, see the *Handbook of Population Census Methods* (3 pamphlets), published by the Bureau of Statistics of the U.N.

[16] This laboratory, directed by Mr. Robert Pagès, holds a cumulative directory of its members' publications at the disposition of researchers (latest edition: 1965). It includes a good number of works dealing with mathematical methods and the use of computors.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

unless the qualitative is given quantitative significance. But up to what point is it legitimate to express in figures complex phenomena involving emotional and sentimental factors? Does not the abstraction which the researcher is obliged to construct threaten to distort the concrete reality which he proposes to study? These are delicate and controversial questions.[17] It would seem reasonable to suppose, however, that the drawbacks of an overly brutal mathematical methodology will be corrected not by abandoning the electronic machine but, on the contrary, by making it work harder, Since it handles figures easily, a larger number of variables can be brought into play, so that the calculations will cover the multiple aspects of a question and thus stick closer to reality.

Care must also be taken to keep the machine from oversimplifying the difficulties of interpretation that statistics always present, whatever their object. When statistics are worked out by hand, the results obtained are figures and percentages which express a relationship between two categories of facts, one inscribed on the "abscissa" and the other on the "ordinate" axis, or among several categories of facts whose curves are superimposed on one graph. But the answer obtained from the calculation is meaningful only if the question has been posed correctly and if the facts treated statistically can reasonably be juxtaposed. The figures may manifest a relationship, but they do not specify whether it is direct or indirect and give nothing in the way of a "causal" interpretation. If, for example, a statistical study of leprosy in relation to rivers and seas is made, the results will show that this disease is more frequent at the water's edge. It would be a mistake to conclude that water is the cause of

[17] Mr. Jacques Ellul, in his book *Propagandes* (Paris, A. Colin, 1962), takes a highly sceptical attitude concerning the validity of mathematical methods in psycho-sociology, and goes so far as to denounce them rather categorically in a paragraph devoted to the ineffectiveness of methods designed to measure the success of a propaganda campaign (Annex I, pp. 294-295 particularly). He notes (p. 286, note 2) that certain American authors contest the premises of public opinion polls, for example Blumer, "Public Opinion and Public Opinion Polling," in the *American Sociological Review*, 1948. On the question as a whole, see: Sorokin, *Fads and Foibles in Modern Sociology and Related Sciences*, Henry Regnery and Co., Chicago, 1956.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

leprosy. Leprosy occurs more frequently in regions bordering on the sea and rivers simply because human agglomerations are more dense there. Its direct "cause" is rather promiscuity and lack of hygiene than anything having to do with water. In the same way,[18] statistics reveal that the number of students at the Faculty of Law in Paris and the number of people with telephones in the Parisian region are closely correlated (*r* close to 0.9). It should not be concluded from this that people with telephones invariably enroll at the Faculty of Law, or that law students invariably have telephones. The augmentation both of students and of telephones is sparked by a third factor, the rise in the average income of Parisians. Statistical correlations must thus be interpreted with extreme prudence, and hasty identifications of causal relationships must be avoided. Statistics involve the same difficulties of interpretation when they are done by machine, and it would be a mistake to suppose that the percentages furnished by a computor are more significant than those worked out by hand.

The machine is highly useful where statistics are concerned because it can handle large quantities of information quickly and easily. It makes it possible to work with a broad range of samples and, as a result, to conduct surveys which give a more faithful image of reality. Using the machine's facilities, the researcher can make a statistical study of the various aspects of a phenomenon and choose intelligently, among all the correlations which the machine proposes, the one which is most likely to translate a direct and causal relationship.

*

A systematic review of all the disciplines in which the electronic machine comes to the aid of researchers would be too lengthy an operation. By way of illustration we will, therefore, examine the field of philology in some detail. In this domain, where not too long ago all the work was done by hand and seemed intractable to automatization, computors already play a

[18] This example is borrowed from Philippe Mouchez, *Démographie* (Collection "Thémis", Paris, Presses Universitaires de France, 1964), p. 134, note 1.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

considerable and ever growing role. Literary studies, in fact, tend increasingly to be founded on exhaustive surveys and precise statistics.[19]

When the electronic machine is given a test in perforated form, it can very easily draw up a lexicon of the author by making a list of all the words he employs with their references and their various forms. Thus the Laboratoire d'Analyse statistique des Langues anciennes at the University of Liège, under the direction of Mr. Delatte and Mr. Evrard, has established a glossary of several works of Seneca. At the University of Tubingen, Dr. Hübner is drawing up a dictionary of Goethe; at Gallarate, near Milan, Father Busa is directing his particular attention to the vocabulary of St. Thomas Aquinas, etc.

Useful though it may be, however, the lexicon of an author furnishes only partial information. The "concordance" is a far more convenient tool, since it notes not only the word (in a certain form) and its reference, but the entire sentence or phrase in which it appears. In this way it is possible to examine the meaning of terms in their immediate context, to compare parallel passages, and to study the thought of an author as well as his style, The "verbal" concordance (which deals with words) was devised by the French Dominicans in the 13th century for use with the Latin Bible. One can imagine the immense amount of work it would take to arrange the words and the phrases in which they appear, not in the order of the text, as was done originally, but in alphabetical order according to their various forms. For the Bible this work would involve making and classifying at least 300,000 cards. Such a task is crushing when done by hand but not terribly difficult when the machine takes over. The technique of "verbal concordance" is thus coming into general use and is being applied to all kinds of authors: Latin classics like Tibullus, theologians like St. Thomas Aquinas, etc.

There is, moreover, a close relationship between a concor-

[19] Details concerning the services which the electronic machine can render in philology can be found in: Dom Jacques Froger, "Emploi de la machine électronique dans les études médiévales," in the *Bulletin de la Société Internatio‑ale pour l'étude de la philosophie médiévale* (Louvain), vol. 3 (1961, pp. 177-188. See also the works cited above, note 6.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

dance and an elaborated glossary or dictionary. The *Littré*, for example, differs from a concordance only in that it gives simply a choice of the characteristic uses of words instead of an exhaustive list; it is a sort of abridged concordance of the French language. Then again, the glossary of a particular author is inevitably monographic in character. A more ambitious undertaking would be to collect a complete corpus of the vocabulary of an entire language. The "*Trésor de la Langue Française*" is being constituted at Nancy with this goal in mind. A large computor, the Gamma 60 (Bull) is devoted exclusively to assembling examples taken from literary works in the form of a concordance. These examples will furnish the citations for a future *Dictionnaire historique national.*

In the domain of lexicography, the machine is capable of all sorts of acrobatics. For example, the "*Trésor de la Langue Française*" is establishing an inverted Littré, "where the words are arranged alphabetically by their last letter instead of the first, a trick which facilitates the study of terminations and inflections. Nothing would prevent an arrangement of words by length, or according to any other principle. Once the machine has registered words, it does whatever it is told to with them.

The lexicographical studies which we have just discussed are accompanied and rounded off by stylistic studies. By putting together the vocabulary of an author and his writing procedures, a picture of his characteristics can be obtained: these are his "habits" in the traditional terminology, but they could also be called his "spectrum," using the word as physicists and chemists do. Such "habits" reveal the personality of a writer just as the rays of light decomposed by a prism indicate the nature and chemical composition of a body brought to the point of incandescence. The electronic machine, by providing more thorough surveys and subtler statistics, makes it possible to define the "habits" or "spectrum" of an author more precisely. It brings to the light traits hidden in the texture of the text, which would undoubtedly be overlooked in a "naked eye" examination; in any event it endows observations with a more objective character.

The use of the electronic machine in stylistic studies raises only one objection, and one which is easy to answer. To study the style and the grammar of an author the text must first be

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

indexed, that is conventional signs must be added, indicating the nature and function of the words as well as all points which the machine will be asked to examine. This is a time-consuming and delicate task and can be handled only by a specialist. Is it worth-while? In answering this question two types of situations must be distinguished: if only a single stylistic element is in question, if for example, one intends to examine only the position of the subject and complement in relation to the verb in order to study inversion, then it would certainly be quicker to pick out by hand the information on which the statistics will be based. The use of the machine, on the other hand, becomes highly advantageous when a very complete index is to be made, covering a quantity of details, which will enable the machine to answer a wide variety of questions, both those with which the original researcher is concerned and those which might interest other philologists in the future.

The kind of lexicographical and stylistic analysis which the machine can do helps philologists in a number of ways. For example, it allows scholars to trace the evolution of a writer throughout his literary career and, as a result, to establish the chronology of his works, if it is uncertain; the order in which they were written is determined, thus giving a relative dating, if there are no fixed chronological points, or even an absolute dating if enough points of reference have been established. An examination of the *Dialogues* of Plato, for example, seemed to show that this great writer increasingly avoided hiatus as, in growing older, he gradually perfected his style. Studies done manually by Lutoslawsky and taken up again on the electronic machine have demonstrated that this was indeed the case; as a result, the *Dialogues* can be arranged in chronological order simply by placing them in the order of decreasing percentage of hiatus. In this way dialogues for which no date was known can be situated in relation to those for which an approximative date has been established. Studies based on analogous principles have been made by P. Guiraud on the chronology of certain of Rimbaud's writings.[20]

[20] Cf. P. Guiraud, *Problèmes et méthodes de la statistique linguistiqu*‹ (Paris, 1960).

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

The "spectrum" of an author can also be helpful in criticism of authenticity, by demonstrating whether or not a particular writing is the work of the author to whom it has been attributed, or by helping to discover the author of an anonymous work. The principle is the following: writings whose "spectrum" is the same can belong to the same author, although this can only be stated as a possibility and not a certitude (since various individuals from the same cultural milieu can have the same characteristics); those whose "spectrum" is clearly different cannot be attributed to the same author, assuming that a writer does not change his habits significantly during his lifetime. Thus Mr. Marichal, who is studying Rabelais at the Laboratoire d'Analyse Lexicographique at Besançon, hopes to settle the question of the authenticity of the Fourth Book. He feels, reasonably enough, that a plagiarist could easily imitate Rabelais' more blatant characteristics, above all his extravagant vocabulary, but could not pay attention to subtle details like word order: the electronic machine will point up the nuances which betray an imitator or reveal the hand of the author. The same principles can thus be applied to proving or disproving a dubious attribution, tracking down long interpolations, etc. By way of example, these resources have been used to study the Epistles of St. Paul, the *Imitation of Christ*, Chaucer;[21] analogous procedures will perhaps shed light on the obscure problem of Shakespeare, which has aroused such quantities of controversy, and on many medieval authors, like Roger Bacon, Albert the Great, to whom an immense body of apocryphal work has been attributed.

Nonetheless, one must be prudent in applying the double principle that two works with the same spectrum can correspond to the same author, whereas different spectrums indicate different authors. Up to what point does an author remain faithful to himself? Within what limits can his spectrum vary according

[21] Cf. G.U. Yule, *The Statistical Study of Literary Vocabulary* (Cambridge, 1944) concerning the author of the *Imitation of Christ*; G. Herdan, "Chaucer's Authorship of the Planetis," in *Language*, 32, 1956, p. 254-259; Rev. K. Graystone and G. Herdan, "The Authorship of the Pastorals in the Light of Statistical Linguistics," in *New Testament Studies*, 6, 1959, pp. 1-15.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

to the periods of his life and the literary genres he tackles? And above all, how sound are the methods employed to sketch the picture which characterizes an author? These are the questions which must be answered. The trouble is that, in the historical disciplines, conclusions are not subject to any control. Physical theories undergo the test of facts; if an engineer makes a mistake in his calculations while constructing a bridge, the bridge caves in; but an historian is not likely to see a medieval author rise from his tomb to contradict him. Shouldn't philologists, and scholars in other fields, test somehow the methods they intend to apply in circumstances in which the results cannot be verified? To be certain that a system of calculation designed for the electronic machine works, the calculation is first done by hand, and then fed into the machine according to a program which is deemed appropriate. If the results do not agree, the machine is wrong, or rather the method imposed on it was worthless. In the same way a philologist, and above all one who deals with criticism of authenticity, should never venture to formulate conclusions on ancient authors without having tried his methods out on modern authors first. This can be done by studying the works of an author who undoubtedly wrote them himself, without the aid of his secretary. If the machine declares that their spectrums are so different that they cannot be the work of one person, then the method is faulty and will produce misleading results when it is applied to ancient authors.[22]

The electronic machine's ability to study an author's habits also serves textual criticism in the conjectural operation on which it often depends. The "historical" or "genealogical" method, when favorable circumstances permit its application, is undoubtedly the best way of restoring the original form of a text

[22] The review *Science et Vie*, No. 572, vol. CVII, May, 1965, contains an item which hides a serious warning behind an amusing facade: "In 1963 a theologian, Father Morton, demonstrated, on the basis of a semantic analysis carried out by an electronic computor, that the fourteen epistles of St. Paul could not all have been written by the same person, and that six different authors were involved. Applying the same analytical methodology to the works published by Father Morton himself, another clergyman, Father Ellison, proved that they could not all have been composed by the author. The electronic brain had 'demonstrated' that, logically speaking, Father Morton did not exist."

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

and eliminating the mistakes of scribes. If the manuscripts can be classified as models and copies, or ancestors and descendants, then the common ancestor of all the existing manuscripts, whether it, is still extant or is reconstructed with the aid of its descendants, will be the one which resembles the original most closely. But only rarely can the genealogical tree be established on the basis of external evidence alone; it is almost always necessary to hunt for internal clues, that is the tenor of the text and its variants. Here the machine plays the preparatory role which we have described above. Comparing the manuscripts, it tracks down the variants and establishes an outline which places them in purely "differential" relationships, disregarding the real structure of their genealogical relationships. Once these concrete operations of calculation have been done, the philologist must do the "intelligent" work, which consists, passing from the the relative to the absolute, in examining mistakes rather than simply differences or variants; this qualitative evaluation is indispensable not only in the method "by groups" which we advocate but also in the method of "common errors" in its classical form and in all critical methods. The copy whose text is most faithful to the original is the one which contains the least mistakes, and the process of distinguishing which reading, among several, is most likely to belong to the author and which are inaccuracies of transmission inevitably involves a certain amount of conjecture. Now the "habits of the author" is one of the most important conjectural criteria, along with the "habits of the scribes" which we will discuss shortly. If two manuscripts give different readings, or forms for the same spot in the text, it is natural to conclude that the one which conforms to the author's usual vocabulary and style is the authentic version, while the one which departs from these habits is the faulty version. The machine facilitates this qualitative evaluation enormously: in providing the philologist with a writer's "spectrum," the machine enables him to judge the reading with objectivity and discrimination. Conjecture is thus built on solid foundations and loses the element of guess-work which rightly attracted criticism when it was founded, as it used to be, on intuition and flair.

Another conjectural criterion often used by philologists is

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

based on the "habits of scribes" or their "psychology." The principle is the following: given two rival readings for the same spot in a text, it is reasonable to conclude that the false one is the one which contains the type of mistakes which scribes generally tend to make. The problem is to pinpoint the "occasions" which lead scribes to make mistakes, and the manner in which they are made. This conjectural procedure, codified (for Latin) by Louis Havet's *Manuel de critique verbale*, has been practised only by hand thus far, and the electronic machine should give it a considerable boost. In most cases, the scribe's mistakes which the "verbal criticism" deals with are inferred on the basis of conjecture, and are not actually identified through the comparison of a copy with its immediate model. Instead of embarking on hazardous speculations concerning the the psychology of ancient and medieval scribes, wouldn't it be wiser to experiment with modern "scribes," who cannot be terribly different from their medieval counterparts? The archives of printing-houses hide a mountain of material, for typesetters are, after all, simply scribes in modern guise. Shouldn't these mines of information be exploited? Here it would be possible to juxtapose the author's text and the first proofs before correction, that is a model and its immediate copy. Hundreds of thousands of copyist's mistakes could thus be directly observed and studied statistically. The electronic machine would be able to pronounce objectively (taking account of the type of printing machine used and the arrangement of its keyboard) on the various types of mistakes, their relative frequency, the occasions which provoke them, etc. I would even measure the extent to which mistakes depend on the personality of the copyist and on his individual psychology. The list of errors most frequently committed by a certain scribe would thus outline his "spectrum," somewhat as stylistic and vocabulary habits reveal that of an author. Founded on this experimental basis, the study of the psychology of ancient scribes would gain in rigor.

Conjectures based on the habits of the author and of scribes must be applied in all cases where only a single manuscript is extant. In these circumstances they are the only way of identifying and correcting mistakes. They are used in identifying or reconstructing the common ancestor, of several manuscripts when

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

the original is missing. They are even more indispensable when a work comes down to us in the form of a single copy, conserved and rediscovered by pure chance. The most important example of this type is the Dead Sea scrolls. These precious documents contain, apart from long fragments of the Hebraic Bible, previously unknown writings which furnish information on the currents of thought in Palestine in an epoch close to that of Christ. Unfortunately, the scrolls resided in jars in the bottom of a grotto for close to 2,000 years; during this period their edges were eaten away, and part of the text is missing. In order to reconstruct the contents of these gaps, Father Busa has used the electronic machine to study the "habits of the author" in the portions of the text which are still legible. The machine does not, as the general public imagines, "automatically" restore the missing pieces, but it does supply an invaluable basis for conjecture.

The lexicographical and stylistic research which we have just discussed are not the only ways in which the electronic machine can be applied in the field of philology. Others could easily be imagined. For example, automatization will undoubtedly come to the aid of the new discipline known as "codicology," to which the review *Scriptorium*, published in Belgium under the direction of Mr. François Masai, is particularly dedicated. Since the codicologist studies ancient manuscripts as archaeological objects, couldn't he handle them as Mr. Gardin treats objects found in archaeological sites? Each manuscript, like a piece of pottery or a flint axe, would be represented by a perforated index card listing its characteristics: material, format, prinpricks, rulings, quartarian signatures, disposition of the text, etc., as well as the place and date when they are mentioned by the scribe or can be ascertained. Mechanical or electronic means could then be used to establish classifications, identify groups by cultural zones, and obtain all kinds of information concerning the history of the book before printing which would help to date and localize manuscripts whose age or origin is uncertain.

The abbreviations found in Latin and Greek manuscripts could also be classified by the machine. Once a relatively complete inventory was built up it could be used to trace the history and origin of certain abbreviations. This would provide

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

additional indications for dating and localizing manuscripts. Projects of this type have been adopted by Mr. Samaran and Mr. Marichal, who are having a methodological survey of Latin abbreviations made at the Institut de Recherches et d'Histoire des Textes; Mr. Glénisson, director of the Institut, even intends to set up a center to study the use of computors in the various branches of learning which work with manuscripts.

The electronic machines have already proved so useful in the literary and philological fields, and offer such promises for the future that the majority of the large universities throughout the world employ them or have acquired one for their own use. Centers of "automatized" philological research have become so numerous that a certain amount of disorder has resulted: it has become apparent that projects are sometimes duplicated by people who are unaware of each other's work. Thus two concordances of the Corpus Tibullianum have just appeared, one published in Italy and the other in the United States, while a third is being prepared by a member of the Laboratoire d'Analyse statistique des Langues anciennes at Liège and will soon appear. In order to avoid such waste of energy and to let researchers profit from each others' methodological experience, the Laboratoire of Liège has just created (at the beginning of 1965) an institution whose French-English title is *Organisation internationale pour l'étude des langues anciennes par ordinateurs—International Organization for Ancient Languages Analysis by Computer.* The seat of the organization is at 2, rue Charles Magnette at Liège. A trimestrial bulletin will report on philological work using electronic machines which has been completed, is being carried out, or is being projected throughout the world, as well as information concerning the methods adopted by the various researchers. By coordinating activities, this institute will give them a fresh impetus. A growing number of philologists from all specialities, particularly certain American Anglicists, have applied for membership in the organization. In view of this interest and of the fact that the methods are the same for modern and ancient languages, Professor L. Delatte is already considering broadening the International Organization of Liège into a world-wide information center for all philological work with computors, without distinction as to language.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

*

One final technique will soon reform, if not revolutionize, the use of the electronic machine: automatic reading.

Up to the present, the machine has been able to read only letters written in accordance with well-defined conventions: their form is a sort of compromise between a completely conventional code and a legible character. The material prepared for the machine is so arranged that it resembles ordinary letters sufficiently to be recognized as such by the naked eye. The intention here is less to have the machine read normal letters than to enable the public to read a code designed more particularly for the machine, and often printed with a magnetic material that stamps directly on the receiving organs of the machine. For several years, large banks have been using checks with number and references written in this way and can thus handle them electronically. The postal services use the same process to handle letters.

Research is currently being carried out with the opposite goal in view: to enable the machine to read letters destined for the human eye and printed in ordinary printer's ink. Professor René de Possel, Director of the Institut Blaise-Pascal of the CNRS (23, rue du Maroc, Paris), is now constructing a machine which achieves this in the following fashion: the space occupied by a line of text is divided ideally into a grid with 72 small squares in each vertical column. A ray of light projected obliquely from the top to the bottom of the papers explores all the little squares in a vertical column one by one and all the columns from left to right. The light diffused by the paper is registered by a photomultiplier at an angle different enough from that which would correspond to Cartesian reflection to avoid the effects of reflections due to printer's ink. The amount of light diffused by the paper varies according to whether the spot on which the ray falls is white or black. The distinction between the parts of the paper where there is writing and those where there is not is translated into the language of the machine, which recognizes only "all or nothing," by signals in the form of the symbols 0 and 1, or "bits," which are then sent to an electronic computor, sometimes by the intermediary of a

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

tape. The information collected in this way is treated according to a program which has the machine "recognize" the letters by equating the symbols formed by a combination of "bits" with an "alphabet" which is placed at its disposition. The characters in this alphabet (from 500 to 1000, for example) include capitals and lower case forms, both roman and italic, of the Latin, Cyrillic, Greek, Gothic, round-hand and English alphabets, plus a number of mathematical and other figures. The machine can thus read any printed page, despite the diversity of characters, since the thousands of letters *a* (for example) of the different fonts used by printing-houses are all translated into the same symbol. By the same process, the machine could apparently identify one figure out of a far larger number, up to a thousand, and could thus apply automatic reading to the Far Eastern alphabets and to handwriting.[23] This project could be useful in various ways, particularly to the blind, and would complement the automatic reader.

The automatic reader currently works at more than 60 letters (or more than a line) per second, a speed which would enable it to read the entire Bible in less than 18 hours. The use of a commercial computor, however, slows down the analysis of results significantly. A special computor, designed to follow the rhythm at which the reading machine explores the text, is now in construction and will speed the process up considerably.

Automatic reading will overcome the handicap which limits the electronic machine most seriously today: the bottleneck at the point of entry. The machine's work is braked by two bottlenecks, one at the entrance, which consists in the time-consuming and often indispensable manual process of preparing the information to be introduced in perforated form, and the other at the exit, where the results must be printed up by some mechanical means. The bottleneck at the exit has now been palliated considerably by improvements in the larger machines which enable them to print or photograph their results at speeds of up to twenty thousand lines per minute. The bottleneck at

[23] Mr. de Possel is currently studying a machine to recite texts as well. This project could be useful in various ways, particularly to the blind and would complement the automatic reader.

the entry remains to be solved, but it will be the day that the reading machine has been perfected: information will then be introduced into the machine practically instantaneously and with no danger of error. It is hard to forecast how long it will be before automatic reading becomes a working reality. The machine which Mr. de Possel is currently constructing should be completed within two years. It will then have to be mass-produced for the general market, so that it could be in current use in five or six years. At this time it will become the indispensable complement of the electronic machine for a whole complex of problems. It should be noted that the two types of machines will aid each other mutually: automatic reading will become increasingly rapid as the electronic computors become more powerful, and the computors, in turn, will function more quickly when they benefit from a system of automatic entry.

The perspectives that automatic reading open are so vast that it is difficult to describe, or even to imagine the services that it will render. To begin with, it will give documentation and automatic translation fresh impetus and jog them out of the state of stagnation into which they have fallen. This is the opinion expressed first by Professor Andreev of Leningrad and then by Mr. Sestier and Mr. Vaucuois at the Centre de traduction automatique of the CNRS in France. It was Professor Andreev's views which persuaded Mr. de Possel to tackle the problem of automatic reading as a whole.

The reading machine will aid philology and linguistics considerably. It will probably not be very useful for ancient and medieval manuscripts, at least in the near future. Those written in uncial Latin or Greek scripts and in Carolingian miniscule will be the easiest to read automatically, since their letters are relatively regular and distinct. But they are, at the same time, the most ancient and important manuscripts, those which philologists are obliged to examine by "naked eye" and with extreme care. The more recent manuscripts (13th to 15th centuries), on the other hand, are extremely plentiful and it would therefore be desirable to have the machine read them quickly. But they are also the type of manuscript which cannot be handled by an automatic reader, since their letters are formed irregularly,

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press

particularly in the cursive script, joined by ligatures and complicated by all types of abbreviation signs. Moreover, whatever their age or handwriting, manuscripts contain corrections made by various hands: erasures, marginal or interlinear additions, words or letters crossed out or emphasized, etc., and all these details, infinitely precious to the philologist, would throw the machine off the track.

Whatever the case may be for ancient manuscripts, the automatic reading of printed material will facilitate the task of preparing critical editions of works which have come down to us only in printed form and of drawing up lexicons and concordances of ancient works on the basis of modern critical editions. The idea of a corpus of tapes and lexicons including all the works which have been written, from antiquity to the present, will lose, at least in a certain measure, its utopian character. In stylistic studies the indexing stage will be immensely simplified: for example, the machine reads a text (printed or typed) and prints it at a rate of one word to the line; in the blank space the philologist marks (in typewriting or careful handwriting) all the information which he considers necessary. The machine rereads the indexed text, and is able to answer any question which the philologist wishes to pose. When automatic translation has been perfected as well, the process will be even simpler. If the machine can translate, this means it is able to understand the meaning and the grammatical function of words. With a stretch of our imaginations, we can thus dream of a process of automatic indexing, analogous to that which researchers still hope to achieve in the field of documentation. Given a printed text, the machine itself would furnish the material necessary for a stylistic study. A single program, established once and for all, could be used to treat any text written in a given language.

Automatic reading, in short, is the process which promises to be the most beneficial, in all the domains which harness electronic resources, and from which the most extraordinary progress is expected.

https://doi.org/10.1177/039219216501305206 Published online by Cambridge University Press