La lecture automatique et l'analyse statistique des textes

Les réflexions que voici ont été provoquées par une conversation avec Madame Fossier, au cours même du Colloque. Nous nous étions posé chacun de notre côté la question suivante : puisque la lecture automatique est attendue dans un avenir prochain, n'est-ce pas perdre son temps, pour le moment, que de réaliser péniblement et lentement des travaux de perforation qui bientôt se feront automatiquement et très vite ?

Pour répondre à cette question, il faut considérer les conditions requises pour utiliser la machine, les services que l'on peut en espérer, et les délais qu'entraîne sa construction pour l'usage courant.

Il ne s'agit évidemment pas de la lecture de caractères imprimés avec des encres magnétiques ou magnétisables, et que l'oeil humain peut reconnaître aussi bien que la machine. Ce procédé est maintenant largement commercialisé, et sert par exemple pour trier automatiquement les chèques bancaires dont les références sont imprimées avec une encre spéciale. Il rend de grands services dans un domaine que l'on peut appeler celui de la comptabilité, mais il est totalement inutile au philologue. La machine dont je parle est celle qui, par des procédés optiques, peut lire des textes écrits avec une encre ordinaire, imprimés, dactylographiés, et même des écritures manuelles.

Ce n'est pas ici le lieu d'entrer dans des détails techniques sur les

différentes méthodes de lecture optique. Disons seulement que la machine doit d'abord enregistrer les signes et pour ainsi dire les récrire à son usage personnel, puis les identifier en disant quelle lettre ou quel chiffre est représenté par tel ou tel signe.

Bien des philologues trop peu informés des procédés de la lecture automatique parlent, comme si c'était quelque chose de très simple, de "donner un manuscrit à lire à la machine". Il ne faut pas oublier que la machine ne peut reconnaître et lire que les signes dont on lui a donné la clé ; on doit mettre à sa disposition, dans sa "mémoire" une sorte de "dictionnaire des signes", de manière que, rencontrant, par exemple, les multiples formes de la lettre *a*, imprimée en les divers types de fonte (capitale, bas de casse, droit, penché, italique, etc.) ou tracée à la main en cursive, elle puisse dire que ces signes variés représentent tous la même lettre *a*.

L'établissement d'un programme suppose donc que l'on a dressé, "à la main" évidemment, un inventaire raisonné de tous les différents signes graphiques que l'on veut faire lire par la machine. Une des conditions essentielles pour qu'un texte puisse être lu automatiquement est (si je ne me trompe, et en tout cas avec certains procédés de lecture optique) que les éléments graphiques soient bien distincts les uns des autres. J'ai fait voir au Professeur de Possel (de l'Institut Blaise Pascal à Paris) des photographies de manuscrits ; pour l'onciale, dit-il, pas de difficulté, car les lettres sont à la fois régulières et séparées ; la minuscule caroline peut encore aller ; mais la machine est impuissante devant les ligatures de courbes du XIIIe siècle, et plus encore devant les cursives des XIVe et XVe siècles où les lettres s'enchaînent inextricablement

les unes aux autres et où l'écriture se complique d'une multitude d'abréviations, elles-mêmes souvent liées aux lettres par un trait de plume ininterrompu. Si l'on ajoute à cela que la machine sera déconcertée par les lettres refaites, les corrections sur la ligne, en interligne ou en marge, ne pourra pas déceler les changements de main et les parties récrites sur grattage, ni relever une foule de particularités archéologiques éminemment instructives pour le philologue, il faut conclure que la lecture automatique, telle du moins qu'on peut l'envisager actuellement, n'est d'aucun secours pour le déchiffrement des manuscrits.

Elle pourra fort bien, en revanche, s'appliquer aux écritures modernes qui se conforment à des canons bien définis, si variés soient-ils, c'est-à-dire à l'écriture imprimée, dactylographiée, ou tracée avec soin à la main. On entrevoit tout de suite les immenses services que l'on peut en attendre, notamment pour l'analyse statistique des textes, puisque le travail se fait précisément soit sur des éditions imprimées, soit sur des copies dactylographiées, soit sur des *listings* fournis par l'ordinateur. Il en va de même pour la confection des index ou tables générales d'immenses répertoires : ceux qui s'adonnent à de telles entreprises assurent que l'investissement le plus lourd, en main d'oeuvre, capitaux et délais, porte sur le travail préparatoire de perforation. C'est la "perforeuse" et sa machine "perforatrice" qui seraient économisées si nous avions à notre disposition la lecture automatique : la machine lirait les textes ou documents de toutes sortes, et ferait elle-même, très vite, les perforations qui aujourd'hui exigent tant de temps et tant d'argent. Ainsi disparaîtrait le goulot d'étranglement qui ralentit à l'heure actuelle le fonctionnement de l'ordinateur.

Combien de temps faudra-t-il encore attendre avant de pouvoir utiliser cette machine ? Où en est maintenant la lecture automatique ?

Elle existe déjà en laboratoire, ce n'est pas douteux : plusieurs électroniciens me l'ont dit. Reste seulement à la commercialiser. Il faudra pour cela résoudre des problèmes techniques et financiers : en effet, la machine doit être assez robuste pour ne pas se détraquer à chaque instant entre les mains d'un utilisateur maladroit ; elle ne doit pas non plus coûter trop cher, pour ne pas exiger une mise de fonds trop importante et s'amortir assez rapidement. Il est difficile de se prononcer sur la date à laquelle la machine fera son apparition sur le marché. Il y a deux ou trois ans, j'avais consulté sur ce point le Professeur de Possel et M. Morton Nadler qui travaillaient à la lecture optique par des méthodes différentes, le premier à l'Institut Blaise Pascal, le second pour Bull General Electric. D'après les prévisions qu'ils m'avaient alors communiquées, par lettre privée ou en conversation, la machine devrait être à la disposition des utilisateurs depuis environ un an et demi : or nous n'avons encore rien vu venir. Les choses n'ont pas été si vite qu'on le pensait. La raison de cette lenteur ne réside probablement pas dans les seules difficultés techniques ; j'ai entendu dire qu'on retardait délibérément la commercialisation de la machine à lire pour éviter de bouleverser l'équilibre économique des entreprises en rendant brusquement inutiles des appareils (par exemple les perforatrices) auxquels il faut laisser le temps de s'amortir. Peut-être, dans ces conditions, attendra-t-on encore quatre ou cinq ans (sinon plus) avant que la lecture automatique n'entre en service.

On revient alors à la question que nous posions en commençant : quelle est la façon la plus rationnelle de mettre à profit les quelques années du prochain avenir ?

Certains se montrent pessimistes à l'égard de la nouvelle invention, et pensent que la machine à lire ne présentera pas une "fiabilité" suffisante. Monsieur Zampolli signale que la lecture automatique est faussée si les caractères sont mal venus dans l'imprimerie ou tracés négligemment à la main, de sorte que le pourcentage d'erreurs sera trop élevé pour qu'on puisse le négliger : sous prétexte d'aller plus vite, la machine introduirait des inexactitudes regrettables dans les analyses statistiques, et il vaut mieux, selon lui, renoncer à en attendre de l'aide.

L'objection a toute sa force si elle s'adresse au lecteur optique mis sur le marché par I.B.M. il y a déjà deux ou trois ans, et qui peut lire les chiffres et cinq lettres, éventuellement tracés à la main. Le démonstrateur qui présentait la machine au SICOB en 1968 recommandait à ceux qui tentaient l'expérience d'écrire avec beaucoup de soin : si les boucles des chiffres 6, 8, 9 et 0 n'étaient pas rigoureusement fermées, la machine se refusait à les lire. Mais il s'agit là d'une machine que l'on peut considérer comme rudimentaire par rapport à celles qu'on nous annonce. Les futures machines sont précisément faites pour s'accommoder de lettres mal venues ou mal tracées, et demanderont à l'utilisateur d'interpréter lui-même les signes qui lui sembleront douteux ou illisibles. Il y a tout lieu de penser que leur "fiabilité" sera sensiblement la même que celle des ordinateurs ; les machines électroniques ne sont pas

complètement à l'abri des erreurs dues à un fonctionnement défectueux, et il suffit d'un grain de poussière sur un disque pour qu'un chiffre soit changé en un autre, une virgule supprimée, etc. : et pourtant on s'en sert sans scrupule.

Une autre observation qui m'a été faite à propos de la lecture automatique est la suivante : "Si on attendait toujours le prochain perfectionnement, on ne ferait jamais rien. Poursuivons donc nos travaux par les procédés actuels sans nous laisser paralyser par la perspective de techniques nouvelles qui viendront un jour remplacer les nôtres".

Cette observation est parfaitement juste en ce qui concerne les progrès techniques imprévisibles, ou que l'on ne peut espérer que dans un avenir très lointain, comme la traduction automatique. Prenons un exemple dans un autre domaine que celui de l'analyse statistique des textes : celui de la restitution critique des mélodies grégoriennes. Les moines de Solesmes qui, à partir de 1840 environ, s'adonnèrent à cette tâche étaient obligés, pour se procurer les manuscrits nécessaires à leur travail, de les copier intégralement à la main. Il fallait des mois pour copier un exemplaire, et des années pour en réunir seulement cinq ou six, ce qui est beaucoup trop peu pour un travail critique approfondi. Les paléographes de l'Abbaye avaient parfaitement conscience de perdre un temps précieux, et s'en plaignaient amèrement. A partir du moment où la photographie a pu s'appliquer aux manuscrits sur une grande échelle, il a suffi de quelques mois pour réunir des centaines d'exemplaires en des reproductions parfaitement fidèles. En 1840,

la photographie existait, mais on ne pouvait guère prévoir qu'elle dispenserait un jour de copier manuellement les manuscrits, et il eût été déraisonnable de rester inactif sous prétexte que les procédés dont on disposait étaient trop lents.

La situation n'est plus tout à fait la même lorsqu'il s'agit de perfectionnements techniques qui, comme la lecture automatique, existent déjà en laboratoire et seront commercialisés sans doute assez prochainement. Notre question initiale se pose toujours : quelle politique faut-il adopter dans le travail d'analyse statistique des textes ? Faut-il poursuivre comme si de rien n'était, ou se croiser les bras en attendant que la machine se charge de la lecture des textes et des perforations ?

La réponse ne peut être que nuancée.

Il ne saurait être question de tout arrêter en prévision de machines qui, après tout, ne seront peut-être dans l'usage courant que dans une dizaine d'années. Jusqu'ici les délais ont toujours dépassé les estimations des constructeurs, et la prudence commande de n'être pas exagérément optimiste quant à la date où la lecture automatique fera son apparition dans le commerce.

D'autre part, il convient cependant de tourner les yeux vers l'avenir, sans limiter notre champ de vision au moment présent. Mieux vaut, semble-t-il, puisque les moyens techniques vont se modifier profondément dans quelques années, ne pas entreprendre des travaux d'une trop grande ampleur, et surtout ne pas prendre de décisions qui engageraient l'avenir, risquant d'obliger à poursuivre

pendant de longues années des travaux commencés par des procédés désormais périmés. Le plus sage parait être, pour le moment, de mettre au point des méthodes et des programmes en cherchant à les affiner le plus possible, de manière à pouvoir se lancer immédiatement dans des entreprises de grande envergure dès que la lecture automatique permettra de les mener à bien beaucoup plus vite et beaucoup plus facilement qu'aujourd'hui.

Mais la seule manière d'élaborer des méthodes, c'est de les expérimenter à mesure, et par conséquent d'exécuter des travaux avec les moyens dont on dispose maintenant. C'est pratiquement ce que l'on fait au L.A.S.L.A. On y a étudié par exemple Sénèque ; les oeuvres de cet écrivain sont des opuscules assez courts, qui par leur brièveté constituent d'excellents terrains d'expérience. Si la perspective de la lecture automatique peut inspirer une ligne de conduite pour les quelques années qui viennent, c'est sans doute de continuer dans la même voie, en s'attachant de préférence aux oeuvres courtes ou à quelques morceaux seulement des oeuvres très longues, en concentrant son effort sur les méthodes et les programmes plutôt que sur l'aspect quantitatif des résultats. Quand viendra le jour où la lecture automatique permettra d'économiser le travail de perforation, les choses iront plus vite et l'on pourra dresser des plans de travail plus ambitieux. En tout état de cause, ce qui est fait, et bien fait, par les moyens actuels, conservera toujours sa valeur.

Dom Jacques FROGER
Abbaye Saint-Pierre
SOLESMES.
(Sarthe)