L'ordinateur et la création littéraire

Communication faite au Colloque International d'Esthétique et des Sciences connexes, Liège, octobre 1966

Mesdames, Messieurs,

Peut-être convient- il d'expliquer brièvement pourquoi et comment un philologue classique s'intéresse à la fois aux ordinateurs et à la création littéraire.

Si la seconde préoccupation va de soi, dès lors que l'on admet une définition large de la philologie, la première paraîtra sans doute plus surprenante. C'est qu'en vérité, beaucoup de chercheurs de ma génération pensent que dans ce domaine fermé que constitue la philologie classique, les études traditionnelles, n'ayant plus de source nouvelle d'information et ressassant depuis des siècles, les mêmes problèmes, ont atteint le point critique où commence la sclérose.

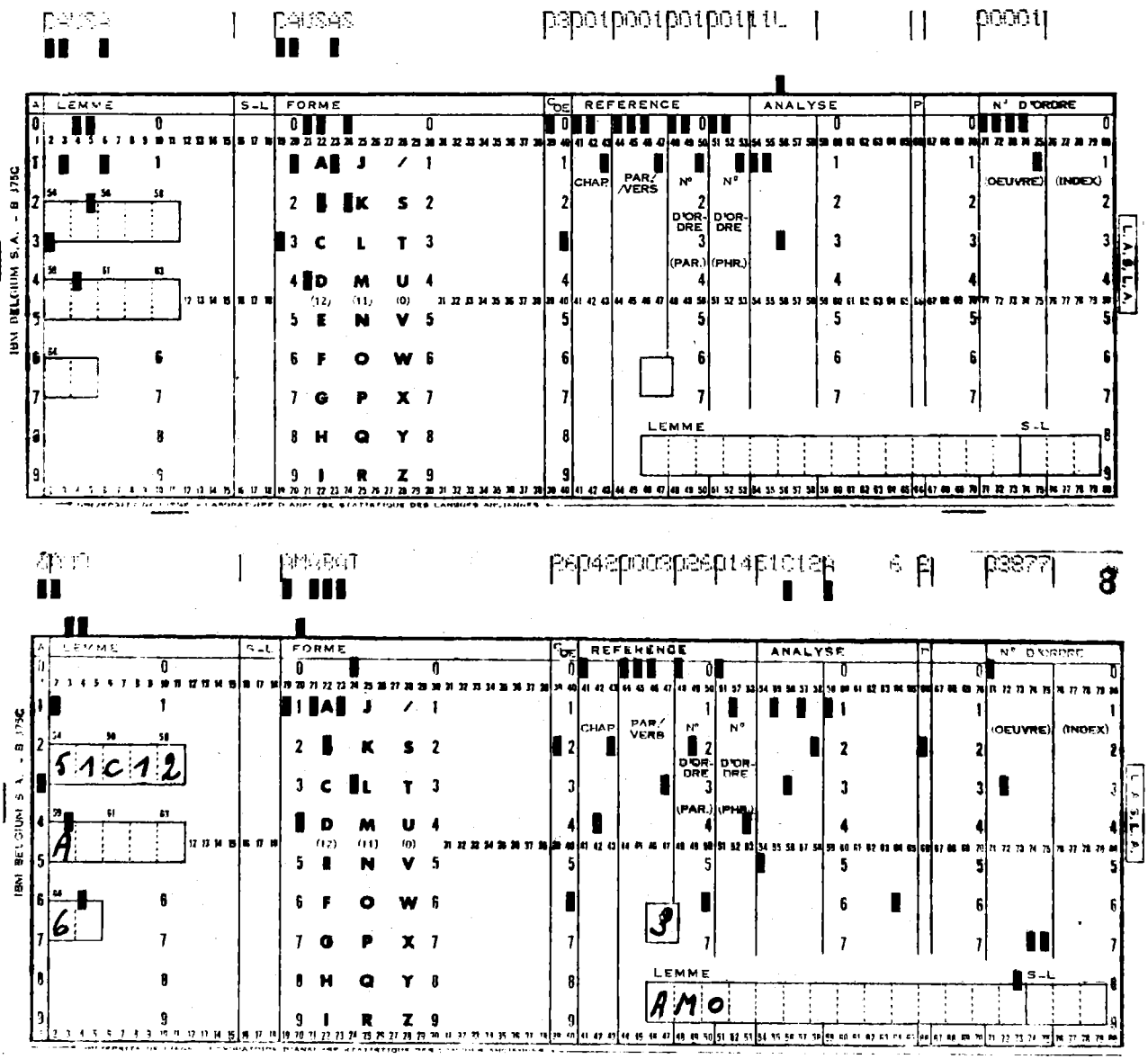
Il m'est apparu que le seul moyen d'apporter un renouvellement à nos études et une solution aux grands problèmes que la philologie classique essaie de résoudre depuis toujours, était de scruter les oeuvres sous leur aspect formel ce qui, auparavant, n'avait jamais pu être réalisé faute de moyens assez puissants.

Mais la technique moderne nous fournit ces moyens: elle met à notre disposition un outil prodigieux, l'ordinateur: une machine possédant un langage et capable entre autres choses -et sans doute est-ce là une définition bien incomplète et bien terne -de faire des opérations logiques et de les combiner de mille manières.

Extrait de la Revue (R.E.L.O.)
II, 1 à 4, 1966. C.I.P.L. - Université de Liège - Tous droits réservés.

Avant d'aborder le sujet propre de cette communication et pour vous rendre plus claire la suite de mon exposé, il me faut bien dire quelques mots de mon Laboratoire et de ce que nous y faisons. Le Laboratoire d'Analyse statistique des Langues anciennes a été créé sur ma proposition en 1961; il est, pour l'heure, équipé de machines mécanographiques classiques: perforatrices, trieuse, tabulatrice, système d'impression automatique et d'un ordinateur IBM 1620 doué de mémoires magnétiques de capacité pratiquement illimitée. Pour apporter à nos études le renouvellement auquel je viens de faire allusion, nous rassemblons actuellement le plus de matériaux possibles. La technique est de mettre sur cartes perforées toutes les oeuvres d'un auteur à raison d'un mot-forme par carte, chaque mot-forme étant accompagné de son lemme, c'est-à-dire de la forme du mot telle qu'elle figure au dictionnaire, d'une analyse morphologique et syntaxique complète ainsi que d'un début d'analyse stylistique qui consiste à indiquer la place de certains mots par rapport au contexte restreint dont ils font partie: la place du complément d'objet direct par rapport au verbe, la place de l'épithète par rapport au mot auquel elle se réfère. Toutes ces analyses sont naturellement symbolisées par des chiffres ou des lettres. Ainsi, les catégories grammaticales sont codifiées par un chiffre en colonne 54, le substantif par 1, l'adjectif par 2, etc. Dans la carte perforée ci-jointe, l'on trouve une forme de mot, un lemme, une analyse.

A partir de cette documentation, nous publions actuellement des Indices accompagnés de relevés statistiques très importants sur la langue de l'auteur étudié. Le but lointain est la solution de problèmes d'authenticité, de chronologie relative que posent certains textes anciens mais aussi naturellement l'étude des thèmes d'inspiration, du style, une description exhaustive de la langue d'un auteur, etc. Ceci nous ramène au sujet de cette communication.



On peut envisager, me semble-t-il, dans la création littéraire, plusieurs plans très différents dont deux sont plus particulièrement intéressants parce que les machines peuvent y intervenir. Le premier est l'aspect purement formel de l'expression au niveau d'une suite de mots: comment un auteur construit-il sa phrase? Comment, dans la suite naturelle du discours se distingue-t-il d'un autre? Par quels éléments formels se différencie-t-il de la norme, de la masse?

Le deuxième aspect est, au niveau de l'idée, celui des thèmes d'inspiration, qu'ils soient conscients ou inconscients. Leur importance est capitale pour qui veut pénétrer profondément dans l'esprit d'un auteur, en comprendre le psychisme, saisir les mécanismes vitaux qui président à la naissance de l'oeuvre littéraire. Un des mérites de la Nouvelle Critique aura été -quoi qu'on en dise -de mettre ce point en évidence.

Dans ces deux domaines de recherches, l'ordinateur et les machines mécanographiques classiques peuvent nous apporter une aide efficace. C'est de cela que j'aimerais dire quelques mots, en me limitant toutefois à suggérer des plans de recherches, à esquisser des lignes de direction.

Tout d'abord, l'aspect formel de l'expression. Une des méthodes d'approche de ce problème me paraît être l'étude des structures profondes de la langue sous leur aspect dynamique de séquence. Comment les mots, sous leur simple aspect de catégorie grammaticale, se répartissent- ils dans la suite naturelle du discours? En d'autres termes, avec quelle fréquence trouve-t-on chez un auteur en suivant naturellement l'ordre des mots dans la phrase, un substantif suivi d'un substantif, un substantif suivi d'un adjectif, un substantif suivi d'un adjectif- pronom, d'un verbe, d'un adverbe, d'une préposition, d'une conjonction de coordination, d'une conjonction de subordination, et d'un point. Puis l'on se pose les mêmes questions à propos de l'adjectif suivi d'un subs-

tantif, suivi d'un adjectif, et ainsi de suite pour chaque catégorie grammaticale à l'exception des numéraux et des interjections dont les fréquences sont trop faibles pour obtenir des résultats intéressants. J'appellerai de telles combinaisons des séquences d'ordre 1. Un petit exemple éclairera la question: dans le vers célèbre "HOMO SUM HUMANI NIHIL A ME ALIENUM PUTO" on trouve les séquences: substantif/verbe, verbe/adjectif, adjectif/adjectif- pronom, adjectif- pronom/préposition, préposition/adjectif- pronom, adjectif- pronom/adjectif, adjectif/verbe, verbe/point. Si l'on néglige le fait qu'il s'agit d'un vers, on aurait pu trouver une toute autre construction telle que par exemple " SUM HOMO A ME NIHIL HUMANI PUTO ALIENUM".

Dans une langue comme le latin, où l'ordre des mots est libre mais non indifférent, on admettra volontiers que de telles listes devraient révéler au niveau le plus formel de l'expression, une manière d'écrire, si l'étude montre -et elle le fait -que certaines structures sont significativement préférentielles ou évitées.

Par ailleurs, ces listes permettent de comparer deux ou plusieurs oeuvres sur le plan instinctif; en effet, elles négligent l'aspect sémantique des textes qui, par définition, se prête mal à une étude comparative et elles ne retiennent que la structure formelle inconsciente des oeuvres.

Maintenant, comment réaliser pratiquement ces listes?

J'ai dit précédemment que nos cartes-mots portaient un code indiquant entre autres choses la catégorie grammaticale du mot. L'ordinateur commence par lire entièrement l'oeuvre et tout en lisant il repère en colonne 54 le code grammatical de chaque mot, il compte les nombres de catégories grammaticales et enregistre dans ses mémoires les 81 combinaisons possibles des séquences d'ordre 1. Lorsqu'il a terminé sa lecture, il se met à écrire, sur sa machine à écrire, les

résultats de ses comptages. Dans le même temps, il calcule et imprime la valeur de chaque séquence si le hasard seul intervenait dans la distribution et enfin, il donne les résultats d'un test statistique qui permet de se rendre compte de la signification des écarts par rapport à cette norme aléatoire. C'est ainsi que,sur l'échantillon ci-joint, vous lisez d'abord le nombre d'observations réelles où l'on découvre les suites d'ordre 1, puis le nombre d'observations théoriques si le hasard seul jouait dans la distribution des structures et enfin la valeur du test calculé entre ces deux types d'observations. Je n'entrerai pas dans le détail du calcul de ce test : il suffit de savoir que dans la dernière colonne toute valeur supérieure à 6,64 indique que la structure observée réellement n'avait qu'une chance sur 100 de se présenter dans le texte ou d'en être absente avec cette fréquence. Tout chiffre au-dessus de 10,83 indique que la structure en question n'avait qu'une chance sur 1.000 de se présenter dans le texte ou d'en être absente avec cette fréquence. On peut donc raisonnablement conclure qu'elle a été voulue ou rejetée par l'auteur et cela avec d'autant plus de force que les chiffres des tests s'écartent de la valeur de 6,64.

Pour une oeuvre de 10.000 mots, la lecture du texte par l'ordinateur, les calculs, l'impression des résultats, prennent une demi-heure de temps. Encore faut-il signaler que le temps de calcul est infime par rapport au temps d'écriture. Etude inconcevable il y a vingt ans seulement!

J'ai dit que l'ordinateur permet aussi de comparer deux oeuvres l'une à l'autre à ce même point de vue.

Lors de l'examen que je viens de décrire, dans le même temps où il imprimait le résultat de ses calculs, l'ordinateur perforait des cartes où il récapitulait les mêmes résultats. Pour comparer deux oeuvres l'une à l'autre, on interclasse ces cartes et l'on part de l'hypothèse selon laquelle les

Echantillon de Séquences d'ordre 1.

	Valeurs absolues	Valeurs théoriques	Valeurs des CHI 2
SUBSTANTIF suivi de			
Substantif	299	378	16,51
Adjectif	135	142	0,34
Adj.-pronom	285	277	0,23
Verbe	511	362	61,32
Adverbe	120	155	7,90
Préposition	79	94	2,39
Conj. coord.	109	94	2,39
Conj. subord.	20	46	14,69
Point	57	70	2,41
ADJECTIF suivi de			
Substantif	264	142	104,81
Adjectif	27	53	12,75
Adj.-pronom	41	104	38,16
Verbe	146	136	0,73
Adverbe	33	58	10,77
Préposition	17	35	9,25
Conj. coord.	62	35	20,82
Conj. subord.	6	17	7,11
Point	19	26	1,88

deux oeuvres sont du même auteur, ou en termes statistiques qu'elles font partie d'une seule et même population. Dans ce cas, chaque oeuvre présente par rapport à l'ensemble des différences qui peuvent être significatives.

Ici encore, l'ordinateur calcule les valeurs théoriques des séquences d'ordre 1, puis les différences par rapport à l'observation et enfin il détermine le même test statistique que précédemment, test qui fait ressortir les structures propres à chaque oeuvre ou chaque auteur.

On peut naturellement se contenter des résultats de l'analyse statistique mais ce serait, je crois, un peu dommage. En vérité, lorsque les structures préférentielles ou évitées ont été déterminées, il faut retourner au texte (un programme d'ordinateur permet d'imprimer sélectivement les deux mots en cause dans une séquence donnée) et tâcher de comprendre pourquoi il y a préférence ou rejet: c'est la tâche du philologue ou du critique.

Voici un petit exemple. J'ai soumis à cet examen les trois consolations de Sénèque, le De Clementia et une oeuvre de Cicéron, le De Senectute. Parmi les séquences qui se sont révélées préférentielles, on trouve dans la Consolation à Helvia et seulement dans cette oeuvre, la séquence: préposition/verbe, qui, à première vue, paraît fort improbable. Après retour au contexte, on constate qu'il s'agit de la construction AD suivi d'un adjectif verbal pour exprimer le but: AD CONSOLANDOS SUOS, AD OBLIGANDA VULNERA, AD MODERANDOS LUCTUS. Ainsi une structure apparaît brusquement dans l'oeuvre de Sénèque et disparaît tout aussi brusquement. C'est donc qu'à un moment donné du temps, le philosophe a considéré cette tournure comme plus élégante, que la subordination par UT, plus expressive, plus philosophique aussi, sans doute en raison du caractère en quelque sorte abstraitement impératif de l'adjectif verbal. Il y a là une sorte de tic du style que l'ordinateur a permis de mettre en évi-

dence. Il y en a bien d'autres. Tous se révéleront utiles lors d'une étude de la chronologie relative des oeuvres de Sénèque.

Je ne me dissimule pas combien cette méthode de recherche des structures est encore primitive et imparfaite et combien le recours au contexte est indispensable dans chaque cas. En effet, une séquence d'ordre 1 est, en latin, très amphibologique: elle peut recouvrir des phénomènes très différents. Ainsi, deux substantifs qui se suivent peuvent être une énumération, mais cela peut être aussi un substantif et un complément déterminatif, un substantif et un objet direct ou indirect, un substantif et son apposé, deux substantifs qui n'ont rien en commun, etc. Malgré les nombreuses constructions auxquelles elle peut correspondre, il est tout de même remarquable de constater que le latin évite toujours cette structure substantif suivi de substantif.

Sans doute, pourrait- on remédier à l'amphibologie de l'analyse en étudiant les séquences d'ordre 2, c'est-à-dire trois mots qui se suivent. L'ordinateur fera le travail avec la même aisance et la même rapidité, mais comme il y a dans ce cas 729 combinaisons au lieu de 81, les données se diluent, deviennent trop faibles pour qu'on puisse leur appliquer un test statistique valable, du moins dans l'état actuel de nos connaissances.

Il importe d'insister sur le fait que la lecture ordinaire d'une oeuvre ne permet absolument pas de se rendre compte de cet aspect de la création littéraire, que l'ordinateur nous apporte, dans ce domaine, une aide incomparable non seulement par la rapidité des opérations mais par la simple détection des séquences significatives.

On peut évidemment objecter que le cas du latin, en raison de l'ordre libre des mots, est un cas privilégié. Trouverait- on des séquences intéressantes en français? Je

crois pouvoir répondre que oui. Sans doute, en français, y a-t-il beaucoup plus qu'en latin, des structures imposées par la langue, mais, lorsqu'on aura fait la part de ces structures, ce qui restera sera d'autant plus significatif.

Un exemple. M. Pierre Guiraud explique, dans Problèmes et Méthodes de la statistique linguistique, que Rimbaud, au cours de sa brève carrière, a renoncé progressivement à la construction substantif/épithète, au bénéfice de la construction substantif/complément déterminatif. S'il avait écrit dans son extrême jeunesse: austérité virginale peu avant d'abandonner la plume, il écrivait: austérité des vierges. Cette différence apparaîtra nettement lors d'une étude de séquences sur ordinateur: on verra diminuer les suites: substantif/adjectif, tandis que les suites substantif/préposition auront une fréquence anormale. Ici encore, et probablement plus facilement qu'en latin, les séquences d'ordre 2 mettraient ces phénomènes mieux en évidence.

Le second problème auquel l'ordinateur peut apporter les éléments d'une solution est celui des thèmes littéraires. La méthode, ici s'inspire de diverses recherches. En particulier, elle prend son point de départ dans la technique des mots-clés.

Je viens de faire allusion à Pierre Guiraud. C'est son nom qui reviendra ici encore. C'est lui qui le premier -et je me plais à lui rendre cet hommage -a proposé d'étudier les mots-clés d'un auteur, ceux-ci étant définis comme les mots de caractérisation qu'emploie un écrivain et qui reviennent sous sa plume avec une fréquence anormalement élevée. Mais Pierre Guiraud n'en a jamais tiré une étude de thèmes. Son but est plutôt une meilleure compréhension des textes d'un auteur par ce qu'il appelle une sorte de lecture verticale, l'emploi d'un mot étant éclairé, enrichi, par les autres emplois de ce même mot. De là, par exemple, l'étude du champ stylistique du mot "ombre", chez Paul Valéry.

Par ailleurs, l'étude des thèmes a fait l'objet ces dernières années de divers travaux, de critiques acerbes et de réponses non moins acerbes. Roland Barthes, Jean-Paul Weber, Serge Doubrovsky, Raymond Picard, que de noms à citer! Encore est -il nécessaire de remarquer que, plus qu'à l'étude des thèmes elle-même, la critique s'adresse à l'outrance des positions de certains néo-critiques, par exemple J. P. Weber, chez qui, finalement tout devient allusion à un thème primordial.

Enfin, il faut bien dire que les recherches de la Nouvelle Critique ne sont pas originales. Les psychiâtres, depuis Jung, se sont penchés sur le problème des thèmes pour tenter de pénétrer profondément dans le psychisme de leur patient, de comprendre ainsi leurs hantises et d'y apporter remède. Ici aussi, la technique n'est pas à l'abri de tout reproche, le plus grand étant le danger pour le psychiâtre de tomber dans la subjectivité.

Ayant ainsi rendu à chacun ce qui lui appartenait, j'en viens à la description d'une méthode de détermination des thèmes qui s'efforce d'atteindre à l'objectivité en automatisant la recherche.

Pour bien comprendre la suite, il me faut faire allusion aux programmes d'ordinateurs que nous avons créés.

L'analyse morphologique de tous les mots d'un texte, dont j'ai parlé précédemment, prend beaucoup de temps lorsqu'elle est faite artisanalement; au surplus, on se fatigue vite à ce travail qui n'est simple qu'en apparence. Notre expérience nous a appris qu'on ne pouvait guère analyser plus de 250 mots par jour: au-delà apparaissent les erreurs de distraction. Tout cela constituait pour nos recherches un sérieux goulot d'étranglement. C'est pourquoi nous avons décidé d'automatiser complètement l'analyse morphologique y compris

la lemmatisation, c'est-à-dire le renvoi de chaque forme du texte à la forme du mot telle qu'elle apparaît dans un dictionnaire.

Actuellement, c'est l'ordinateur qui, confronté avec la forme, nous dit, pour prendre un exemple très simple, que AMABAT est la 3e personne du singulier de l'imparfait actif d'un verbe de la première conjugaison et que la forme de ce mot au dictionnaire est AMO.

Une telle analyse suppose non seulement un programme extrêmement complexe -vous vous en doutez bien -mais encore trois lexiques que l'ordinateur possède dans ses mémoires et qu'il consulte à tout instant: un lexique de toutes les désinences de la langue, un lexique de toutes les formes irrégulières et enfin un dictionnaire général des mots où ceux-ci sont rangés par ordre alphabétique et accompagnés d'un chiffre indiquant leur fréquence dans la langue normale.

Les deux premiers lexiques ne nous intéressent guère pour notre propos actuel puisque le premier n'est fait que de désinences et que les formes du second sont, par définition même, les plus fréquentes et par conséquent les plus communes: elles ne peuvent donner aucune indication sur des thèmes.

C'est évidemment le dictionnaire général qui est l'outil de base pour notre recherche.

Comment ce dictionnaire est-il constitué? La capacité énorme des mémoires de l'ordinateur autorisait plusieurs solutions. L'une aurait consisté à copier un dictionnaire tel que le Forcellini, puisque nous pouvions envisager de mettre en mémoire 40 à 50.000 mots différents. Mais c'eût été là une solution brutale qui présentait certains inconvénients, en particulier, celui de ralentir le rythme de l'analyse. J'ai résolu alors de ne mettre en mémoire qu'une sorte de vocabu-

laire de base, constitué d'abord par les mots des oeuvres qui avaient été analysées avant que nous n'ayons l'ordinateur, c'est-à-dire les trois Consolations et le De Clementia de Sénèque, deux livres de Lucrèce, le Corpus Tibullianum puis par les mots d'un vocabulaire de base latin établi en 1939 par un précurseur de l'étude des langues, l'Américain Diederich, vocabulaire établi sur un dépouillement de divers textes représentant environ 100.000 mots. Tout cela ne fait, chose curieuse, qu'un vocabulaire de 6.000 mots différents seulement grâce auquel nous comprenons plus de 99 pour cent de tous les textes.

Lorsque l'ordinateur a fait l'analyse de tous les mots d'une oeuvre, et lorsque cette analyse a été vérifiée par les philologues, on fait relire l'oeuvre par l'ordinateur en lui donnant un nouveau programme qui lui permet de regrouper tous les mots différents du texte en ordre de fréquence décroissante. Pour arriver à ce but, il compte et enregistre en mémoire, le nombre d'occurrences de chaque mot différent du texte. A la fin de son travail, il imprime et perfore une liste de lemmes en commençant par le mot qui compte le plus grand nombre d'occurrences et en affectant chaque mot d'un chiffre indiquant le nombre de fois qu'il se rencontre dans le texte.

A partir de cette liste, il est très facile de demander à l'ordinateur la liste des mots-clés, c'est-à-dire des mots dont la fréquence est anormalement élevée par rapport à un état moyen de la langue.

Si vous voulez vous souvenir que chaque mot dans le dictionnaire de base de l'ordinateur possède un chiffre qui indique sa fréquence moyenne dans la langue, une comparaison, pour chaque mot, entre cette fréquence et la fréquence dans l'oeuvre étudiée, comparaison accompagnée d'un test statistique quelconque, permet de mettre en évidence les mots-clés.

Ceux-ci sont à la base d'une recherche de thème.

L'on pourrait se limiter à étudier seulement ces mots dans les contextes où ils apparaissent. Mais une telle étude ne serait pas complète. En effet, il est certain que le mot fait partie, dans notre lexique mental, d'un groupe qui exprime la même idée ou une idée analogue. La cohésion du groupe est assurée par des liens affectifs ou intellectuels. C'est pourquoi il paraît difficile d'étudier des thèmes littéraires ou philosophiques sans faire appel à tous les mots qui expriment une idée ou un sentiment analogue à celui qui a été mis automatiquement en évidence par l'ordinateur.

Il faut donc délimiter des champs notionnels dont les mots appartiennent à la même sphère de signification. C'est la dernière tâche préliminaire à l'étude des thèmes: elle aussi peut être automatisée, ce qui garantit l'objectivité de la recherche. Dans ce but, un quatrième dictionnaire est constitué où les principaux mots pleins de la langue sont groupés en réseaux associatifs ou champs sémantiques. Le programme de la machine est constitué de telle manière qu'à la lecture de l'oeuvre, dès que l'ordinateur rencontre un des mots de la famille, il relève automatiquement tous les passages où se trouvent les autres mots de la même famille. Ce travail se fait avec une extrême rapidité puisque l'accès à un mot quelconque dans une oeuvre de plusieurs milliers de mots demande 1/4 de seconde.

Pour illustrer cette recherche et l'aide qu'apporte l'ordinateur, je citerai un exemple. Dans la liste des mots-clés du poète élégiaque latin Properce, on trouve le mot AQUA qui n'est pas particulièrement frappant, puisque les philologues qui se sont occupés de l'oeuvre de Properce n'ont jamais constaté chez cet auteur, un thème de l'eau. Mais, lorsque l'ordinateur, groupant les mots qui expriment la même idée (UNDA, LYMPHA, LIQUOR, MARE, FLUMEN, etc) relève tous les

passages où l'on trouve ces mots, on s'aperçoit qu'il y a chez Properce un thème extraordinaire, une obsession de l'imagination matérielle qui se traduit par des images où l'eau intervient continuellement et dans les contextes les plus divers. L'ordinateur dénombre ainsi plus de 300 occurrences de cette sphère de pensée privilégiée qui révèle un poète marqué profondément par un élément dont les diverses formes conditionnent chez lui un thème obsessionnel toujours présent à l'arrière-plan de sa conscience.

Telles sont, Mesdames, Messieurs, les deux types de recherches où l'ordinateur peut nous apporter une aide précieuse. En terminant cette trop longue communication, je voudrais ajouter deux choses.

La première est que les recherches de ce type ne sont pas nécessairement gratuites. Les psychiâtres -j'y ai fait allusion précédemment et mon savant collègue le Dr Bobon est un pionnier en cette matière -s'intéressent de plus en plus au langage pour des raisons évidentes, soit de diagnostic précoce, soit de pénétration en profondeur de la conscience. Les deux recherches que je viens de décrire sont intéressantes respectivement dans ces deux domaines: les premières pour un diagnostic, les secondes pour une thérapeutique. Il est évident que l'ordinateur, parce qu'il fournit des éléments de réponse avec une extrême rapidité, parce qu'il permet de vérifier dans des temps très brefs, de multiples hypothèses de travail, est, dans ce domaine aussi de la recherche médicale, un outil très précieux.

La seconde observation porte précisément sur cette notion d'outil. L'ordinateur n'est que cela. Toutes les recherches, tous les calculs, dont je viens de parler, pourraient être faits manuellement, artisanalement, sans recourir aux machines. L'ordinateur n'ajoute rien qualitativement à l'homme: il lui donne seulement une nouvelle dimension d'ordre quantitatif et si l'ordinateur fournit les éléments d'une répon-

se, il reste tout de même à l'homme l'interprétation et la décision finales.

Louis Delatte.