L'ETUDE DES TEXTES LATINS AU LASLA

S. GOOVAERTS, Première Assistante à l'Université de Liège

Il y a dix ans que l'Université de Liège, sur la proposition du Professeur L. Delatte, créait le Laboratoire d'Analyse statistique des Langues anciennes (L.A.S.L.A.). Ce Centre était le premier à appliquer les techniques automatiques à l'étude des textes littéraires anciens. On comprend que l'intrusion des machines dans un domaine qui semble peu accessible aux progrès de la technique puisse déconcerter certains esprits. Il est pourtant indéniable que les méthodes traditionnelles ont conduit la recherche philologique dans une impasse et que, depuis plusieurs années déjà, un besoin de renouvellement se faisait sentir en ce domaine. Il est apparu que les techniques automatiques étaient précisément de nature à apporter au chercheur ce renouvellement, d'une part, en lui fournissant une aide précieuse pour l'application des méthodes traditionnelles, d'autre part, en lui ouvrant l'accès à des voies nouvelles d'investigation.

Pour examiner de façon vraiment scientifique la plupart des questions qui se posent à lui, - et en particulier celles où intervient un élément de comparaison, tels les problèmes de chronologie et d'authenticité, - le philologue doit pouvoir disposer de relevés exhaustifs portant sur les divers éléments de la langue. En l'absence de tels relevés, il en est réduit à se contenter de sondages et à fonder ses conclusions sur des impressions subjectives. Or, pour beaucoup d'auteurs anciens, il n'existe même pas un *index verborum* complet. La raison en est évidente. La constitution artisanale de ces relevés exigerait des délais considérables tout en comportant un risque important d'erreurs et d'omissions. Les machines peuvent supprimer ces obstacles. En effet, elles sont capables de restituer fidèlement les données qu'on leur a confiées et d'effectuer avec rapidité et exactitude toute espèce de calcul et de dénombrement.

D'autre part, depuis quelques années, la statistique s'est introduite dans l'étude du langage et du style. Divers procédés d'analyse quantitative, fondés sur le dénombrement des caractères formels des textes, ont été mis au point et ainsi s'est constituée une discipline nouvelle, la statistique littéraire. Il va de soi que cette méthode exige des dénombrements importants et des calculs fort longs et qu'il n'est guère possible de l'aborder sans recourir à l'aide des machines.

La base de tous nos travaux est le fichier mécanographique. Comme unité d'enregistrement des textes, nous avons choisi le mot, élément auquel se rattachent le plus aisément les diverses informations morphologiques, syntaxiques et stylistiques de l'analyse.

La constitution de notre fichier comprend plusieurs étapes.

La première étape consiste à perforer le texte à raison d'un mot par carte. Chaque carte porte la forme telle qu'elle se présente dans l'édition de base, ainsi qu'un code symbolique représentant l'oeuvre traitée et, éventuellement, un code relatif à la ponctuation (fin de phrase, de paragraphe, de chapitre).

La perforation du texte ne requiert aucune connaissance du latin et peut être le fait d'une simple dactylo. La tâche du philologue est de "prééditer" le texte, c'est-à-dire de séparer les enclitiques, de signaler la ponctuation et, une fois le fichier perforé et imprimé automatiquement, de le vérifier soigneusement.

A ce stade, un simple classement alphabétique des cartes permettrait déjà d'obtenir des *indices* et des concordances par formes. Mais ceux-ci seraient de consultation peu commode : en effet, les différentes formes d'un même mot, telles *sumus, est, futurus,* se trouveraient dispersées tandis que seraient confondues les formes homographes de mots différents telles *consuli,* datif de *consul* et *consuli,* infinitif présent passif de *consulo.* Aussi avons-nous jugé indispensable de rattacher chaque forme à un lemme, c'est-à-dire au mot tel qu'il apparaît au dictionnaire : par ex., nous regroupons des formes telles que *sumus, est, futurus* sous le lemme *svm.* Cette opération, théoriquement fort simple, pose en pratique de nombreux problèmes, par ex., la distinction des lemmes homographes (*eo,* adverbe et *eo,* verbe) et le traitement des mots qui changent de catégorie grammaticale (adjectifs et participes substantivés). En ce domaine, il est impossible d'éviter une part d'arbitraire, mais l'essentiel est de s'en tenir strictement aux conventions établies, afin de conserver une uniformité à travers toutes les oeuvres étudiées. Aussi avons-nous choisi un dictionnaire de référence dont nous nous écartons le moins possible.

Une fois lemmatisée, la forme reçoit une analyse très détaillée, exprimée en code; chaque indication morphologique et syntaxique est représentée par une perforation dans une colonne donnée. Pour un verbe, par ex., nous notons la conjugaison, le mode, le temps, la voix, la personne, l'appartenance à une proposition principale ou subordonnée, le mode de subordination.

La lemmatisation et l'analyse du texte se font en une seule étape. Lors de la création du L.A.S.L.A., ces opérations étaient réalisées artisanalement par le philologue. On comprend sans peine qu'elles constituaient pour celui-ci une tâche écrasante et que ce procédé ralentissait nos travaux dans des proportions considérables. En effet, il était impossible d'analyser plus de 250 mots par jour sans risquer des erreurs dues à la fatigue ou à la distraction. Il importait avant tout d'automatiser cette étape.

Depuis 1964, le L.A.S.L.A. est doté d'un ordinateur IBM 1620 équipé de mémoires à disques, ce qui a permis la mise au point d'un programme d'analyse automatique du latin.

Le programme commence par calculer pour chaque mot, à partir des codes symboliques de ponctuation, une référence comprenant le numéro du chapitre, le numéro du paragraphe, le numéro d'ordre du mot

dans le paragraphe, le numéro d'ordre dans la phrase et enfin le numéro d'ordre dans l'oeuvre étudiée.

Ce système de référenciation nous met à l'abri de tout accident en ce domaine : erreur de référence, carte égarée ou déclassée.

Ensuite, pour chaque forme qui lui est proposée, l'ordinateur recherche tous les lemmes et toutes les analyses théoriquement possibles à partir de 2 lexiques et d'une table de désinences enregistrés préalablement sur disques.

Le premier lexique contient les mots invariables et les formes irrégulières avec leur lemme et leur analyse. Le second donne, pour les mots réguliers, le radical, le lemme et un début d'analyse grammaticale.

La table des désinences fournit pour chaque désinence l'analyse correspondante.

Le procédé d'analyse est le suivant : l'ordinateur compare d'abord la forme avec celles du lexique I. Ensuite, il décompose en un radical et une désinence possible, en isolant d'abord la dernière lettre, puis les deux dernières et ainsi de suite.

Il consulte chaque fois la table des désinences et le lexique II pour déterminer si les deux éléments qu'il a ainsi isolés correspondent à une désinence et à un radical existants et pour vérifier l'identité entre l'analyse de la désinence et celle du radical.

L'ordinateur ne peut évidemment lever les nombreuses amphibologies qui caractérisent la langue latine. Aussi perfore-t-il une carte chaque fois qu'il découvre une analyse possible. Le fichier ainsi produit contient donc pour chaque mot plusieurs cartes. Il est listé et c'est le philologue qui choisit, parmi les analyses proposées, l'analyse correcte qui seule sera conservée.

Bien qu'elle ne soit pas entièrement automatique, cette analyse en ordinateur facilite grandement le travail. La tâche du philologue est allégée dans une mesure considérable. Bon nombre d'erreurs dues à la fatigue ou à la distraction sont éliminées et les délais réduits dans une mesure appréciable.

Depuis le 11 février 1971, le L.A.S.L.A. dispose d'un ordinateur plus moderne, un IBM 360/20 et nous travaillons également sur un 360/25, ce qui a exigé une reconversion de nos programmes. La conception générale du programme d'analyse demeure la même. Les modifications intervenues, outre celles qui répondent aux exigences particulières de la machine, visent surtout à simplifier et à accélérer le processus d'analyse.

Une fois le lemme et l'analyse établis, le philologue ajoute les indications qu'on ne peut obtenir automatiquement : indications syntaxiques, renseignements relatifs à la tradition manuscrite.

Le fichier ainsi constitué se prête à de nombreuses exploitations. Celles-ci reposent principalement sur des tris et des dénombrements opérés selon divers critères (lemmes, analyses grammaticales, codes de subordination, etc...). Ces opérations peuvent être réalisées au moyen de la trieuse classique IBM 108 dont dispose le L.A.S.L.A. ou grâce à l'ordinateur. Le recours à ce dernier est indispensable lorsqu'il s'agit de fichiers de grande dimension. Dans ce cas, le fichier est enregistré sur disques, ce qui permet de gagner du temps et de réduire l'intervention humaine au strict minimum.

Nos publications réunissent les principaux documents qu'un tel fichier permet d'obtenir.

Nous publions d'abord un *Index Verborum* où les lemmes, précédés de leur fréquence, sont classés dans l'ordre alphabétique. Sous chaque lemme sont regroupées les formes, dans l'ordre grammatical et, à l'intérieur d'une même analyse, dans l'ordre du texte. Chaque forme est accompagnée de sa référence. Les formes qui posent un problème d'apparat critique sont rejetées en fin d'article.

Vient ensuite une liste des lemmes rangés dans l'ordre de fréquence décroissante, un tableau de distribution du vocabulaire, une liste des 50 mots significatifs les plus fréquents; une série de relevés donne la répartition des occurrences en catégorie grammaticale et pour chaque catégorie en sous-catégorie (déclinaisons pour les substantifs, voix-modes-temps pour les verbes ..).

Pour ce qui est de la syntaxe, nous publions une liste où les verbes subordonnés sont rangés selon le mode, le mot régissant et le temps. Le contenu de cette liste est repris dans un tableau sous forme de chiffres.

Enfin, nous présentons la liste des mots situés en fin de phrase, accompagnés de leur référence et du nombre de mots contenus dans la phrase, ainsi qu'un tableau de distribution des longueurs de phrases.

Une telle publication est loin d'épuiser toutes les possibilités offertes par nos fichiers.

Il serait fastidieux de décrire en détail les opérations mécanographiques qui permettent d'obtenir ces divers documents, mais il convient d'insister sur le fait que nos ouvrages sont réalisés de manière presque entièrement automatique.

Une fois le fichier chargé sur disques, un programme permet la mise en ordre alphabétique et grammatical (ordre d'index). Un programme distinct calcule la fréquence de chaque lemme et produit un fichier récapitulatif qui permettra la constitution de la liste de fréquence.

Un autre programme réalise les comptages grammaticaux et imprime les résultats en valeurs absolues et en pourcentages.

Enfin, l'ordinateur établit la mise en page de l'index, de la liste de fréquence et de la liste des verbes sub-

ordonnés et perfore un fichier destiné à l'impression. A partir de ce fichier, une document-writing IBM 870 imprime automatiquement les documents définitifs qui seront reproduits en off-set. Ce système présente de sérieux avantages sur les procédés traditionnels d'impression qui risqueraient fort de réintroduire des erreurs dans un texte se composant principalement de chiffres.

Le premier auteur que nous avons choisi comme terrain d'expérience pour l'application de nos méthodes est Sénèque le Philosophe. Notre choix a été motivé par le fait qu'il n'existait pour cet auteur aucun *index verborum,* alors que son oeuvre joue un rôle de première importance dans l'histoire des idées et pose des problèmes de chronologie auxquels la critique traditionnelle n'a pu apporter de solution satisfaisante.

A l'heure actuelle, nous avons publié les volumes relatifs aux oeuvres suivantes : les trois *Consolations,* le *De Constantia Sapientis,* le *De Brevitate Vitae,* le *De Clementia.* Les *Epîtres* à Lucilius sont sous presse. Un membre de l'équipe du L.A.S.L.A. prépare une thèse de Doctorat sur l'ensemble des *Tragédies* dont le fichier est achevé.

Enfin les fichiers du *De Tranquillitate Animi* et du *De Otio* sont prêts et pourraient être publiés sous peu.

En vue de recherches d'ordre littéraire, nous avons analysé le *Corpus Tibullianum* et les 4 livres des *Odes* d'Horace. De nombreux étudiants ont réalisé leur mémoire de licence à l'aide de nos machines et nous accueillons chaque année un certain nombre de chercheurs étrangers.

D'autre part, il nous a paru que l'enseignement du latin pourrait être complètement rénové si l'on disposait d'un vocabulaire de base enfin rationnel et d'une grammaire où les phénomènes morphologiques et syntaxiques seraient hiérarchisés en fonction de leur fréquence d'apparition dans les textes. Dans ce but, nous avons entrepris, en collaboration avec l'inspection de l'Enseignement Secondaire, le dépouillement systématique des auteurs étudiés pendant les humanités.

Nous avons publié, jusqu'à présent, les *indices* et relevés relatifs au *Bellum Gallicum* de César, aux *Bucoliques* de Virgile, à la *Conjuratio Catilinae* de Salluste, ainsi que des extraits des *Fables* de Phèdre et des *Métamorphoses* d'Ovide. En outre, le *Bellum Jugurthinum* de Salluste est actuellement sous presse et nous entamons l'analyse des *Géorgiques* de Virgile.

Notre tâche essentielle a été jusqu'à présent de rassembler la documentation étendue que requiert toute étude statistique. A l'avenir, tout en poursuivant les divers groupes de travaux que je viens de mentionner, nous comptons nous consacrer plus particulièrement à l'exploitation de nos données selon les méthodes d'analyse quantitative.

D'une manière générale, ces méthodes permettent de distinguer chez un auteur ou dans une oeuvre donnée les traits qui relèvent de l'usage courant de la langue des traits qui sont propres à cet auteur ou à cette oeuvre. Elles contribuent donc à définir le style en prenant en considération les critères les plus variés (vocabulaire, grammaire, expressions, métrique, phonétique). Dès lors, elles peuvent être appliquées à tous les domaines de la critique philologique et littéraire.

Je me limiterai ici à deux exemples d'application de ces méthodes.

Le premier concerne l'étude du caractère d'un auteur et des mécanismes psychologiques qui ont contribué à l'élaboration de son oeuvre. Lorsqu'on examine la liste de fréquence du vocabulaire d'un auteur donné, on constate la présence de mots dont la fréquence est anormalement élevée par rapport à l'usage courant de la langue. Il est légitime de supposer que de tels mots, dont la répétition est en partie inconsciente, sont de nature à éclairer la personnalité profonde de l'écrivain. Deux études du Professeur L. Delatte, l'une portant sur Properce et Tibulle, l'autre sur les *Odes* d'Horace, illustrent l'intérêt de telles recherches.

Le second exemple concerne le domaine de la critique d'érudition (problèmes d'authenticité, d'attribution, de chronologie). Si l'analyse quantitative ne peut donner une réponse directe à ces questions, elle est de nature à fournir au chercheur une base solide et objective sur laquelle fonder ses conclusions. En effet, toutes ces questions peuvent être envisagées comme un problème de comparaison de style et, sur ce point, à la suite de tests mathématiques portant sur les critères stylistiques les plus divers, l'analyse quantitative permet de conclure à l'homogénéité ou a l'hétérogénéité du style.

C'est dans cette perspective que j'ai étudié les problèmes d'attribution que pose le livre III du *Corpus Tibullianum.*

Ces deux exemples illustrent le type de recherche auquel nous comptons nous consacrer pendant les années à venir.

Je voudrais, pour terminer, souligner que tous nos travaux sont le fruit d'une collaboration étroite entre les divers membres de l'équipe du L.A.S.L.A. Etant donné la complexité de telles études, il est presque impossible à un chercheur isolé de les réaliser. Aussi nous a-t-il paru souhaitable d'étendre cette collaboration.

C'est dans ce but que le Professeur L. Delatte a créé l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur dont la *Revue* informe les chercheurs des progrès de la technique et des diverses études entreprises dans ce domaine.