L'AUTOMATISATION DE L'ETUDE DES TEXTES GRECS AU LASLA

E. EVRARD, Professeur ordinaire à l'Université de Liège

Je ne referai pas ici l'historique du Laboratoire d'Analyse statistique des Langues anciennes (L.A.S.L.A.) du Professeur L. Delatte : M. J. Denooz vous a dit hier tout ce qu'il faut sur ce sujet. Je rappellerai seulement qu'au cours de ses dix années d'existence, le L.A.S.L.A. a consacré la majeure partie de son activité au latin, et spécialement à Sénèque le Philosophe.

Toutefois, dès le début, nous avons manifesté l'intention de faire place au grec dans nos travaux : la dénomination même du L.A.S.L.A., qui fait référence aux langues anciennes, le montre à suffisance. Comme pour le latin, nous avons élaboré pour le grec des méthodes de traitement automatique. La seule différence, c'est que, dans ce domaine, nos réalisations sont nettement moins nombreuses.

Au demeurant, le traitement décrit à propos du latin peut, si l'on y apporte les modifications appropriées, s'adapter aux autres langues et c'est lui que nous avons utilisé pour le grec. Je ne répéterai donc pas en détail ce que vient de vous dire Mlle S. Govaerts.

Comme les textes latins, les textes grecs sont perforés manuellement, à raison d'un mot par carte, avec l'indication des critères de référenciation (fin de phrase, fin de paragraphe, fin de chapitre, éventuellement fin de ligne). Ils subissent ensuite une référenciation automatique, que l'on peut conformer aux usages fixés par la tradition pour les différents auteurs.

Par après, chaque mot est spécifié par un lemme et par une analyse morphologique et syntaxique. L'analyse est exprimée par le moyen d'un code qui s'inspire des mêmes principes que le code latin, mais qui a été adapté aux particularités de la langue grecque.

Ce code a été constitué par M. J. De Bie, qui en a donné une description dans *Revue,* 1966, 4, pp. 31-50 et 1967, 1, pp. 17-38. Notons toutefois que, pour notre usage propre, nous avons quelque peu simplifié le système proposé par M. De Bie.

Enfin, l'exploitation du fichier consiste, pour le grec comme pour le latin, dans la préparation automatique d'*indices,* de relevés statistiques et, si le besoin s'en fait sentir, de concordances. Jusqu'ici, donc, nulle différence.

Tout irait bien si le grec ne présentait une particularité spécifique qui, à toutes les étapes du traitement, fait apparaître des problèmes variés et complexes. Cette particularité réside dans l'accentuation et dans les autres signes adventices qu'utilise le grec : esprits, iota souscrit, coronis, tréma.
Voyons d'abord les contraintes qu'entraîne cette particularité aux deux étapes extrêmes, *input* et *output* : en effet, ces deux étapes, au point de vue des traitements, sont en étroite corrélation.

Le problème qui se pose est celui de la représentation des différents caractères par des codes perforés compréhensibles pour l'ordinateur et susceptibles de commander une impression automatique. La chose est simple lorsque le nombre de caractères ne dépasse pas quelques dizaines. Mais les multiples combinaisons possibles des lettres de l'alphabet grec avec les signes accessoires nous conduisent bien au-delà de cette limite. Voyons ce point en détail.

Les dix-sept consonnes ne font pas de difficulté, sinon que le ρ apparaît avec ou sans esprit rude, selon qu'il est ou non à l'initiale. Mais la situation est toute différente pour les sept voyelles. Celles-ci peuvent se rencontrer soit sans aucun signe accessoire, soit avec l'un des deux esprits, soit avec un accent aigu ou grave, soit encore avec l'un des esprits et l'un des accents. De plus, les voyelles α, ω, ι et υ peuvent aussi être marquées de l'accent circonflexe, combiné ou non avec un esprit. Enfin, α, η et ω peuvent, dans chacune des combinaisons énumérées jusqu'ici, avoir en outre un iota souscrit. De plus, l'usage du tréma fait encore apparaître une série de combinaisons supplémentaires. Le compte de toutes ces possibilités dépasse largement la centaine. Si l'on y ajoute les chiffres, indispensables pour la référenciation, et quelques signes de ponctuation, on arrive aux environs de 150, et ce nombre est dépassé si l'on veut, en outre, réserver des codes spéciaux pour les signes critiques utilisés en papyrologie.
Ajoutons qu'il paraît difficile de renoncer à la distinction entre majuscules et minuscules pour les caractères alphabétiques.

La solution idéale consisterait à réserver à chaque signe, simple ou complexe, une représentation codée occupant une position dans la mémoire de l'ordinateur. Il y aurait, par exemple, un code distinct pour η, un autre pour η accompagné de l'esprit doux, un troisième pour η avec esprit rude, et ainsi de suite. Malheureusement, cette solution excède les possibilités techniques actuelles.

Sans doute pourrait-on croire que, moyennant quelques concessions, elle serait praticable à condition que l'on exploite au maximum les possibilités offertes par le byte de 8 bits qui est utilisé pour la représentation des caractères dans des machines telles que les ordinateurs IBM de la série 360. Ce byte, en effet, admet 256 configurations distinctes et, dans les opérations internes, l'ordinateur est capable de les distinguer toutes. En outre, il est possible d'introduire à volonté n'importe laquelle d'entre elles dans la mémoire centrale, par exemple au moyen de cartes perforées.
Mais, en fait, seule une partie de ces possibilités est réellement exploitable : ce sont les codes auxquels correspond, en standard, une représentation graphique, un caractère d'impression. Pour les autres, il faut, dans l'état actuel du *hardware,* utiliser des procédés lents et difficiles, tels que la technique du

multiple punch en carte perforée. Ceci s'explique, d'ailleurs. Quelque moyen que l'on emploie pour communiquer avec la mémoire centrale, il y a toujours un moment où l'opérateur doit utiliser un clavier et l'on imagine difficilement un clavier qui permettrait 256 frappes distinctes. Le même problème se pose aussi pour l'*output.* En théorie, rien n'empêche d'attacher à chacune des 256 configurations une représentation graphique. Mais, que l'*output* soit une imprimante à barre ou à chaîne, ou encore une machine à écrire automatique, le traitement serait singulièrement ralenti dans ces conditions.

Au demeurant, pour autant que je sache, le *hardware* que l'on trouve actuellement sur le marché ne permet pas l'exploitation pratique des possibilités que je viens d'exposer, et il le permettait bien moins encore à l'époque où nous avons commencé à nous intéresser au problème du grec. Il fallait donc trouver une autre solution.

La méthode que nous avons choisie consiste à traiter séparément le caractère alphabétique, les groupes formés par l'accent et/ou l'esprit, et, enfin, le iota souscrit, le tréma et les signes diacritiques (tels le point, sous la lettre, que l'on emploie pour marquer les lettres de lecture incertaine).

A chacun de ces caractères ou groupes de caractères, nous faisons correspondre un code défini par une combinaison de perforations. Pour les lettres majuscules et pour quelques signes accessoires, ce code est précédé d'un autre code dont le rôle, à l'*output,* est de provoquer le positionnement de la machine en *upper case.* Lors de la perforation d'un texte, les codes des signes accessoires précèdent celui du caractère alphabétique.

Comme *output,* nous utilisons une *Document writing IBM 870* composée d'un lecteur-perforateur de cartes et d'une machine à écrire connectée. Cet ensemble a dû subir une série de modifications pour répondre à nos besoins. La principale concerne le fait que, pour les signes accessoires, il fallait supprimer l'avance du chariot (ou, ce qui revient au même, le déplacement de la tête d'impression). Le système d'impression que nous utilisons ainsi est relativement lent, mais il donne de forts bons résultats, en tout point comparables à ceux que procure la typographie classique.

On pourrait imaginer une technique analogue pour l'imprimante d'un ordinateur : chaque ligne serait imprimée en trois fois : une première pour les accents et les esprits, une seconde pour les autres signes accessoires, une troisième pour les caractères alphabétiques; seule la dernière impression serait suivie d'un échappement vertical du papier. Ce système, pour être praticable, exigerait la fabrication de barres d'impression spéciales.

Il n'est pas utile de donner plus de détails ici sur le système d'*input/output* que nous pratiquons pour le grec. Le lecteur désireux d'informations plus complètes les trouvera dans *Revue,* 1966, 3, pp. 21-45.

Indépendamment des problèmes d'entrée et de sortie, la présence des accents et des esprits fait appa-

raître des difficultés lors des traitements internes. Le cas le plus clair est celui des tris. L'ordre alphabétique ne dépend évidemment que des lettres, à l'exclusion des blancs et des signes accessoires. Dès lors, si l'on opère par des procédés automatiques, le critère de tri, aligné sur la gauche, doit être formé d'une suite ininterrompue de lettres, sans insertion d'aucun autre caractère. Mais, en revanche, ce que l'on veut retrouver en *output* après un tri, ce sont les formes complètes, avec esprit et accent. Dès lors, l'opération de tri n'est possible que si l'on associe à chaque mot son image réduite.

Dans une première étape, alors que nous opérions au moyen d'une trieuse, nous commencions par produire en ordinateur un fichier des formes réduites que nous obtenions à partir du fichier des formes pleines. Les deux fichiers étaient alors interclassés puis un tri avec entraînement sur une trieuse IBM 108 nous permettait d'opérer un classement au cours duquel les deux cartes constituant une paire (forme réduite et forme pleine) restaient liées et étaient ordonnées en fonction des formes réduites. Ce système était lent et fastidieux en raison des importantes manipulations de fichier qu'il exigeait. Aussi avons-nous cherché un allègement à ce procédé.

Cet allègement, nous l'avons trouvé grâce au tri en ordinateur. En effet, en même temps que l'ordinateur prend en charge les formes à trier, il crée de façon toute provisoire les formes réduites qui doivent servir de critère. Ces dernières, après avoir rempli leur office, disparaissent et ne laissent aucune trace dans l'*output*. C'est là un procédé beaucoup plus élégant et plus rapide que celui que nous utilisions antérieurement. Quant à l'effort supplémentaire de programmation qu'il exige, il est vraiment minime eu égard au résultat.

Venons-en maintenant à la lemmatisation et à l'analyse. Comme pour le latin, ce furent là, au départ, deux opérations manuelles. Mais, comme pour le latin, encore que plus tard, il nous a semblé intéressant de mettre au point un programme de lemmatisation et d'analyse automatiques. Ce programme a été élaboré pour notre IBM 1620 et a fait l'objet d'un exposé à la Société des Etudes Grecques à Paris en février 1970. En fait, il n'a guère été utilisé. Son transfert sur le 360/20 dont nous disposons maintenant fait partie de nos projets mais n'a pas encore été réalisé.

Dans les grandes lignes, ce programme procède comme son homologue latin. Pour les mots invariables et les formes anomales, il se contente de consulter un lexique et, en cas d'identité, de prendre le lemme et l'analyse qu'il y trouve. Quant aux formes des mots à flexion, il les décompose en un radical et une finale, en donnant successivement à ces deux éléments les diverses longueurs possibles. A chaque étape, il cherche à identifier le radical en consultant le lexique. En cas de réponse favorable, il vérifie si la finale est une désinence possible pour la catégorie grammaticale dont relève le radical identifié.

Ici encore, l'accent pose quelques problèmes spécifiques.

Tout d'abord, mais c'est là une difficulté mineure, il convient, avant toute tentative d'analyse, d'éliminer

purement et simplement les accents dus à l'enclise et de faire disparaître l'effet de la barytonèse.

Par ailleurs, les variations de l'accent au cours de la flexion empêchent, si l'on n'y prend garde, l'identification des radicaux. Il est donc indispensable, tant dans le lexique et les tables de désinences que dans les mots à analyser, de traiter la forme réduite, accompagnée d'un code indiquant la nature et la place de l'accent. L'analyse comporte alors, après les étapes décrites ci-dessus, une opération supplémentaire de contrôle de l'accent. Dans le cas des indéclinables, cette opération vérifie simplement l'identité du code d'accent dans le lexique et pour le mot à analyser. Quant aux mots fléchis, la tâche est un peu plus complexe. Elle consiste à déterminer si l'accent du mot est compatible avec l'accent premier dont le code est fourni par le lexique.

Les fichiers de textes grecs préparés suivant les techniques que je viens de décrire se prêtent aux mêmes exploitations que celles qui ont été exposées à propos du latin. Je ne m'y attarderai pas et signalerai simplement la récente parution d'un index grec préparé automatiquement au L.A.S.L.A. Il se trouve dans l'édition de Basile de Césarée, *Sur l'origine de l'homme* par A. Smets et M. Van Esbroeck (Coll. Sources chrétiennes, n° 160).

Ce qui, en revanche, me paraît utile, c'est de mentionner ici les travaux entamés au L.A.S.L.A. pour l'exploitation automatique des papyrus documentaires grecs. On sait que le traitement de ces textes présente des particularités liées surtout à l'état dans lequel ils nous sont parvenus. Avec l'assistance de deux papyrologues, MM. J. Bingen et A. Tomsin, nous avons, en 1968, perforé sur cartes une série de papyrus, en veillant à enregistrer tous les signes diacritiques relatifs à la tradition textuelle. Nous avons ensuite établi des cartes consacrées aux corrections proposées par divers philologues.

A partir du fichier ainsi constitué, nous avons réalisé divers types d'éditions : édition princeps; édition intégrant les corrections; transcription diplomatique, ne reproduisant que les éléments réellement lus sur le document et négligeant les conjectures et corrections d'éditeurs. Nous y avons joint un index, un index des noms propres, une liste de fréquence des lemmes, un index inverse et des modèles de concordances sélectives.

Je me bornerai à quelques réflexions sur celles de ces réalisations qui ne se trouvent pas habituellement dans nos publications relatives à Sénèque. L'index inverse peut rendre de grands services aux éditeurs de textes qui s'efforcent de combler par conjecture les lacunes de la tradition manuscrite. Il arrive, en effet, que seule la fin d'un mot subsiste et que le début en soit perdu. De ce point de vue, l'index inverse est particulièrement utile aux papyrologues, qui traitent des textes souvent fort abîmés. Quant aux concordances sélectives, elles sont de divers types. La plus simple fournit, pour un mot donné, la référence des passages où il se trouve ainsi que, pour chacun d'entre eux, un contexte dont l'utilisateur fixe la longueur à son gré. Le critère de sélection peut être constitué non par un mot mais par un groupe de lettres consécutives, appartenant ou non à un même mot. Enfin, un type de concordance à sélection

complexe fournit les contextes qui contiennent deux ou plusieurs mots donnés. Il n'est pas nécessaire, me semble-t-il, d'insister sur l'aide que trouveraient les philologues s'ils pouvaient interroger les textes par le moyen de telles concordances sélectives. A chaque fois, en effet, il leur serait loisible d'adapter au mieux leurs questions à leurs besoins.

L'ensemble des réalisations que je viens d'énumérer a fait l'objet d'une publication que nous avons présentée au 12e Congrès de papyrologie (Ann Arbor, août 1968). Elle a paru sous le titre *Choix de Papyrus grecs. Essai de traitement automatique* par J. Bingen, A. Tomsin, A. Bodson, J. Denooz, J.C. Dupont, Et. Evrard (Travaux publiés par le L.A.S.L.A.), Liège, 1968.

Depuis cette publication, le L.A.S.L.A. a continué ses recherches dans le domaine de la papyrologie : MM. A. Tomsin et J. Denooz se préparent à exposer les résultats nouveaux qu'ils ont obtenus au Congrès qui se tiendra cette année à Marburg.
Telles sont, brièvement caractérisées, les réalisations du L.A.S.L.A. dans le domaine du grec. Mais, avant de terminer mon exposé, j'aimerais de vous soumettre quelques réflexions fondées sur une expérience de dix ans concernant l'utilisation des ordinateurs dans les sciences humaines.

En premier lieu, nous sommes persuadés que les chercheurs qui désirent utiliser l'ordinateur doivent acquérir la formation technique grâce à laquelle ils pourront réaliser eux-mêmes les tâches d'analyse et de programmation. Cette méthode est beaucoup plus efficace que celle qui repose sur la collaboration d'un philologue et d'un technicien. Dans ce dernier cas, en effet, le fait que chacun d'entre eux ignore le domaine de l'autre provoque inévitablement des malentendus, des pertes d'information, des gauchissements.

Ma seconde remarque portera sur l'utilisation de codes symboliques, tels ceux que nous employons pour l'analyse morphologique. Un code satisfaisant doit tendre à un équilibre entre deux exigences contradictoires : il devrait être à la fois simple et exhaustif. S'il est trop simple, il manque d'efficacité. Mais, pour être exhaustif, il risque de se compliquer au point d'en devenir impraticable. Au reste, il ne faut pas céder trop facilement au mirage de la codification. Souvent, on imagine qu'elle est une sorte de panacée, alors que, dans bien des cas, l'enregistrement des données sous leur forme naturelle constitue la méthode à la fois la plus simple et la plus efficace.

Enfin, il faut observer que, dans le domaine des ordinateurs, les progrès techniques sont d'une rapidité telle que les utilisateurs ont peine à les suivre. Il en résulte un danger : c'est que, trop souvent, le *software* est en retard par rapport au *hardware*. Dans bien des applications prévues pour les ordinateurs les plus récents, on remarque des particularités qui paraissent liées aux contraintes d'un matériel antérieur. Ici, comme en tant d'autres domaines, il faut constamment remettre en question les idées et les habitudes.

Quoi qu'il en soit, les travaux que nous avons engagés depuis dix ans au L.A.S.L.A. nous ont donné la

ferme conviction que, désormais, l'ordinateur et les techniques automatiques sont des auxiliaires indispensables de la recherche, même dans les domaines auxquels, de prime abord, ils paraissent le moins adaptés.