LE LEL DE LIEGE

EXPOSE PAR L. DELATTE

Il nous a paru nécessaire de répartir en trois exposés, la description des lexiques du L.A.S.L.A. que nous allons avoir l'honneur de faire devant vous. Peut-être convient-il de dire aussi dès le départ - mais cela paraîtra évident - combien tout le travail qui va vous être exposé est une affaire d'équipe et d'une équipe parfaitement homogène: nous avons la chance de former une telle équipe que vous pourrez juger sur les résultats de ses travaux.

Je commencerai par un exposé général.M. Evrard parlera de la technique et de l'exploitation du dictionnaire de mots lui-même et M. Bodson traitera des lexiques de désinences et de préverbes que comporte notre programme d'analyse du latin.

L'exposé général comprendra deux parties. La première sera consacrée à un historique qui définira les buts du L.A.S.L.A. et les méthodes mises en oeuvre pour y atteindre. Cet historique est nécessaire parce qu'on ne peut comprendre notre programme d'analyse lexicale et morphologique qu'à la lumière des buts que nous souhaitons atteindre et parce que c'est la lente maturation de nos méthodes de travail qui a permis l'élaboration d'une symbolique et finalment d'un programme qui, je l'espère, vous séduira. La seconde partie de l'exposé, consacrée au programme lui-même est indispensable pour situer parfaitement la place des lexiques dans leur cadre général et pour préciser les exigences auxquelles ils doivent répondre, puisque les lexiques nous sont utiles dans la mesure même où ils nous permettent de lemmatiser les textes et surtout d'analyser la morphologie.

Avant de commencer, il me reste un agréable devoir à remplir. Je me plais à redire ici la dette de reconnaissance que nous avons contractée envers le magistral précurseur, le Père BUSA.

C'est, en effet, la vitalité de son propre Centre dont on parlait déjà dans toute l'Europe quand nous n'existions pas encore, qui nous a encouragés à nous engager dans la voie où nous sommes maintenant. Que le Père Busa veuille bien me permettre de lui dire notre profonde admiration pour l'oeu-

vre de pionnier qu'il a accomplie.

Les buts que nous nous proposions en créant en 1961 le L.A.S.L.A. n'étaient cependant pas les mêmes que ceux du Centre de Gallarate. La raison profonde pour laquelle le Laboratoire a vu le jour, qui explique et justifie l'orientation prise par nos travaux, était l'espoir que nous avions -et que nous avons toujours- de résoudre à l'aide de techniques nouvelles, les grands problèmes auxquels la philologie classique se heurte depuis toujours et qu'elle a été impuissante à résoudre: je veux parler des problèmes d'authenticité (telle oeuvre a-t-elle été écrite par tel auteur?), les problèmes de chronologie relative (tel dialogue de Sénèque a-t-il été écrit avant tel autre?) et, enfin, les problèmes des mécanismes de la création littéraire, qu'ils soient conscients ou inconscients.

Nous pensions -et nous pensons toujours- que l'analyse des textes, sous leur aspect de structures formelles, sans référence à leur valeur sémantique, apportera des éléments de réponse qui auront une valeur infiniment plus objective que les arguments traditionnels de la philologie classique dont on connait la subjectivité et partant la réversibilité.

Une telle conception du travail postulait évidemment que nous ne nous limitions pas à faire des *Indices* ou des Concordances mais exigeait que nous réunissions à propos de chaque mot du texte ancien, le plus grand nombre possible de renseignements. La grande hantise a été de ne rien oublier puisque, une fois le travail entamé, il serait bien difficile de revenir en arrière.

Après d'inévitables hésitations, nous avons résolu de faire une analyse morphologique et syntaxique complète de chaque mot du texte ainsi qu'un début d'analyse stylistique formelle qui consiste à étudier l'ordre des mots dans la phrase.

Voilà donc avec quels espoirs et quelles prétentions nous avons commencé nos travaux. Et, je le répète, ce sont les buts que nous nous sommes assignés qui ont conditionné toute la technique et toute la méthodologie du Laboratoire et, en particulier, le dessin et le contenu de la carte perforée tels que je vais les décrire maintenant.

La base de tous nos travaux, c'est l'oeuvre littéraire antique qui se présente à nous sous l'aspect formel d'une suite définie de mots. C'est donc chacun de ces mots que nous allons traduire en langage-machine, en prenant comme principe de n'écrire qu'un seul mot par carte. Vous trouverez sur la carte 1, la place où la dactylo perfore la forme du mot telle qu'elle apparaît dans le texte: dans les colonnes 19 à 38. Mais comme ces cartes-mots seront

mélangées à la suite de divers tris ou par accident de manipulation, comme d'autre part, le nombre des cartes va atteindre très vite d'énormes proportions, il est essentiel que chaque mot soit accompagné de références extrêmement précises qui permettront de le retrouver et de le reclasser facilement. Chaque mot est donc situé (c'est ce qui apparaît sur la carte 2) d'abord par un code d'oeuvre en colonnes 39 et 40: 04 est le code de la Consolation à Marcia, puis plus précisément par le numéro du chapitre, le numéro du paragraphe, le numéro d'ordre dans le paragraphe, le numéro d'ordre dans la phrase et enfin par son numéro d'ordre dans l'oeuvre entière. Ces références occupent les colonnes 41 à 53; puis 71 à 75. Les colonnes 76 à 80 sont réservées au numéro d'ordre du mot dans l'index et ne sont perforées naturellement que lorsque l'oeuvre est prête à l'impression.

Ces références sont perforées automatiquement dans chaque carte par une petite calculatrice IBM 602 A. On n'imagine pas, en effet, qu'une opératrice soit obligée de perforer chaque référence à propos de chaque mot: ce travail conduirait à de multiples erreurs. La dactylo se contente de perforer, dans la colonne 66 un code qui, par convention, signifie fin de chapitre, fin de paragraphe ou de vers, fin de phrase. La machine reconnaît ces codes comme étant des ordres d'exécution sélectifs: chaque fois, par exemple, qu'elle détecte une perforation indiquant une fin de chapitre, elle augmente son compteur de chapitres d'une unité cependant que, dans le même temps, elle remet à l'unité ses compteurs de paragraphes, de numéro d'ordre dans le paragraphe, de numéro d'ordre dans la phrase.

Ce système de références, outre qu'il nous est utile pour l'étude statistique des longueurs de phrases, nous met à l'abri de tout accident puisqu'un contrôle de séquence fait sur la trieuse ou sur l'ordinateur signale que l'ordre est perturbé ou qu'un élément de la série est manquant.

Jusqu'à présent, nous n'avons encore sur la carte que la référence et la forme du mot telle qu'elle apparaît dans l'oeuvre. Or certaines formes d'un mot sont si différentes les unes des autres que la mise en ordre alphabétique serait impossible, si on ne les regroupait pas toutes sous un même lemme, c'est-à-dire sous la forme du mot tel qu'il figure au dictionnaire. Comment, par exemple, regrouper les différentes formes de *esse (sum, eram, fui),* si l'on n'a pas le lemme *sum* comme référence générale?

A défaut d'une telle organisation, une liste de mots-formes est inutilisable, pour qui veut faire des décomptes sérieux. Il faut donc que nous transcrivions sur la carte le lemme de la forme: pour un verbe, ce sera la première

personne du singulier de l'indicatif présent actif, pour un substantif, la forme du nominatif singulier. Comme dictionnaire de référence, nous avons choisi, en raison de sa cohérence, le *Lexicon totius latinitatis* de Forcellini.

La carte porte maintenant la forme du mot, le lemme et la référence: c'est le cas de la carte numéro 3.

Une nouvelle difficulté se présente ici: c'est le problème des homographes. Si l'on se réfère à un lemme sans autre indication plus précise, beaucoup de mots et de formes s'écrivant de la même façon, mais n'ayant pas le même sens ou la même construction, se confondent dans les listes. Après divers tâtonnements, nous avons décidé d'affecter les lemmes homographes d'un indice numérique. Au départ, cet indice était conditionné par l'ordre dans lequel les lemmes homographes apparaissent chez Forcellini. Nous avons dû renoncer à cette règle pour des impératifs techniques d'analyse en ordinateur. Actuellement, l'ordre adopté pour l'attribution des indices de lemme correspond à l'ordre de succession des catégories et sous-catégories grammaticales dans notre codification numérique des mots. Lorsque deux lemmes homographes se présentent, nous attribuons l'indice 1 à celui qui se rangerait le premier dans cette succession. Ainsi *eo* (verbe) recevra l'indice 1 (les verbes sont codifiés 5) et *eo* (adverbe) recevra l'indice 2 (les adverbes sont codifiés 6). Lorsque deux lemmes homographes appartiennent à la même catégorie grammaticale, le plus fréquent -selon notre expérience- reçoit l'indice 1: ainsi *tempus 1*, le temps et *tempus 2*, la tempe.

Cette question de lemmes réglée, il reste l'analyse morphologique et syntaxique. Nous avons été obligés naturellement de codifier ces analyses dans un langage alphanumérique très condensé pour utiliser au mieux les quelques colonnes qui restaient disponibles sur la carte.

Vous trouverez dans le grand tableau en annexe, la liste des diverses indications, ainsi que le numéro des colonnes qui sont réservées à ces renseignements.

Colonnes 54 à 65:

1° Distinction des catégories grammaticales en 54

2° Analyse détaillée:

Substantifs: déclinaison, cas, nombre.

Adjectifs: classes, cas, nombre, degrés de comparaison, emplois substantivés ou adverbiaux, genre. Vous remarquerez que, pour condenser les informations, nous avons été amenés à utiliser une codification alphabétique dans un tableau à double entrée: le zo-

ning ayant une signification et les digits en ayant une autre. Ainsi, un adjectif de la première classe étant codifié par 1 et le comparatif par le zoning 12, la résultante est évidemment un A.

Numéraux: type, éventuellement, cas, nombre, genre.

Adjectifspronoms: type, cas, nombre, distinction des emplois pronominaux, adjectivés et adverbiaux, genre.

Verbes: conjugaison, mode, temps, voix, nombre et personne; caractère périphrastique des formes; distinction des verbes de propositions principales et subordonnées (pour ces derniers, indication codée de leur mode de subordination); place de la proposition subordonnée par rapport à la principale. Pour les participes, gérondifs et adjectifs verbaux, certaines de ces indications sont remplacées par celles du cas et du nombre.

Adverbes: degré de comparaison, nature grammaticale.

Prépositions: cas régi; notation des expressions du type *mecum*.

Conjonctions:distinction des conjonctions de coordination et de subordination (pour ces dernières, notation du temps et du mode régi; cette notation est mentionnée aussi sur les cartes des pronoms et adverbes relatifs et interrogatifs).

Pour être complet, il faut ajouter, quoique cela sorte un peu de sujet de ce Séminaire, qu'en colonne 65 on trouve des codes relatifs à la tradition manuscrite:

1 désigne une leçon contestée
2 un mot absent dans certains manuscrits mais adopté par l'éditeur
3 une conjecture reçue
4 une lacune comblée
5 une *crux*
6 une leçon rejetée par l'éditeur mais compréhensible
7 un mot présent dans certains manuscrits et rejeté par l'éditeur.

Tel a donc été le travail d'analyse et de codification réalisé artisanalement par les quatre membres de l'équipe du L.A.S.L.A. pendant les quatre premières années de son existence. Nous disposions alors d'un matériel IBM purement mécanographique: une perforatrice 26, une trieuse statistique 108, une petite calculatrice 602 A, un système d'impression automatique 870 et une tabulatrice 448.

Lorsque la dactylo avait perforé dans la carte, la forme du mot telle

qu'elle apparaît dans le texte, et lorsque les cartes avaient été référenciées sur la 602 A, chaque membre de l'équipe prenait son fichier à sa table de travail et, à la main, écrivait sur chaque carte, dans des emplacements prévus à l'avance pour des commodités techniques, le lemme et l'analyse complète du mot.

On se rend compte aisément de la somme énorme de travail que représentent ces analyses. Elles exigent une attention sans défaillance et une connaissance parfaite du texte, jusque dans les plus petits détails; elles ne souffrent pas la moindre approximation. C'est la raison pour laquelle un philologue ne peut guère analyser que 200 à 250 mots par jour. Au surplus, nous nous sommes astreints à de nombreuses vérifications qui se font à deux niveaux. D'abord à un niveau simplement matériel: il faut contrôler s'il n'y a pas d'erreur de frappe de la part de l'opératrice; au niveau philologique ensuite: les membres de l'équipe se réunissent et examinent ensemble les diverses analyses qui pourraient prêter à discussion.

Enfin, au dernier stade du travail, la dactylo perfore dans les cartes les annotations de l'analyse. Après une dernière vérification matérielle, nos cartes sont prêtes à être utilisées. Voici, par exemple une carte complètement rédigée: carte 4.

De 1961 à 1965, soit dans un espace de 4 ans, nous avons publié 5 volumes consacrés aux 3 Consolations de Sénèque, au *De Clementia* et au *Corpus Tibullianum*.

Pendant toutes ces années, désespérés par la lenteur, la lourdeur de l'analyse, nous avons cherché des moyens d'accélérer le travail. L'un de ces moyens a été d'automatiser, pour les mots indéclinables et qui ne peuvent prêter à confusion parce qu'ils n'ont qu'une signification possible, la lemmatisation et l'analyse morphologique: c'est le cas, par esemple, pour certaines prepositions: *ab, de, ex*.

Puis, il nous est venu à l'esprit que l'on pourrait aller beaucoup plus loin si l'on disposait d'un ordinateur.

J'ai donc demandé à notre Fonds National de la Recherche Fondamentale les crédits nécessaires à la location d'un petit ordinateur scientifique IBM 1620 avec une mémoire centrale de 20.000 positions seulement,mais équipé de deux puissantes mémoires périphériques constituées chacune par une armoire à disques magnétiques, IBM 1311. On peut enregistrer sur chaque pile de disques, deux millions de caractères numériques ou un million de caractères alphabétiques, répartis en vingt mille secteurs qui ont chacun leur adres-

se propre et qui forment en fait 100 cylindres concentriques. La lecture et l'écriture sur disque se font naturellement par des têtes magnétiques placées comme au bout d'un peigne qui effleurent la face intérieure des disques par un déplacement latéral.

L'utilisation des disques combine la consultation séquentielle et la consultation aléatoire: toute lecture et toute écriture sur disque commence à un secteur quelconque désigné, dans le programme, par son adresse et se continue en conséquence pour un nombre de secteurs déterminé lui aussi par le programme; la consultation peut évidemment être limitée à un seul secteur.

Le positionnement des têtes de lecture sur le cylindre désiré demande en moyenne 25 millisecondes; encore faut-il ajouter que si deux lectures successives concernent un même cylindre, aucun positionnement nouveau n'est nécessaire. Les opérations d'entrée et de sortie (lecture et perforation des cartes, lecture et enregistrement sur disque) sont de quelques millisecondes. Quant aux opérations internes (opérations arithmétiques, operations logiques et transferts), leur temps est de quelques micro-secondes.

L'ordinateur est arrivé en janvier 1965 et après avoir suivi des cours de programmation et nous être familiarisés avec le maniement de la machine, nous avons réalisé en huit mois un programme d'analyse automatique de la morphologie du latin, qui fonctionne parfaitement depuis près de deux ans.

Ici, nous commençons à toucher de plus près le but de cette réunion. Vous voudrez bien excuser la longueur du préambule, mais l'historique du L.A.S.L.A. me paraissait nécessaire pour faire comprendre comment nous sommes arrivés à codifier puis à automatiser les informations linguistiques.

Avant que MM. Evrard et Bodson ne prennent la parole et pour situer parfaitement leurs exposés, il me paraît interessant de décrire brièvement notre programme. En très gros, ce programme consiste:

1. à comparer chaque mot des textes à un dictionnaire de formes irrégulières et de formes invariables;
2. à dissocier, dans chaque mot du texte, la racine du mot de sa désinence;
3. à comparer la désinence ainsi mise en évidence à une table générale des désinences mises en mémoire;
4. à retrouver la racine du mot dans un dictionnaire général, mis, lui aussi, en mémoire;
5. pour les formes des verbes composés, à séparer les préverbes du verbe.

Voici deux exemples qui permettent de mieux comprendre le travail de

la machine. Soit la forme invariable *ad*. On ne peut évidemment la décomposer. Il faut donc l'analyser en un seul bloc et pour cela disposer d'un lexique où les formes sont reprises telles quelles, accompagnées d'un code indiquant leur valeur morphologique. Le cas est le même pour toutes les formes des mots variables mais irréguliers, pour lesquels la décomposition en radical et désinence est possible, mais inefficace. L'ordinateur compare la forme du texte à la forme qu'il trouve dans son lexique et sort une carte sur laquelle il a perforé ce qu'il a lu dans le lexique.

Soit maintenant une forme verbale régulière telle que *amabat*. On y distingue deux éléments: l'un, *am*, commun á toute la flexion du verbe, l'autre, *-abat*, essentiellement variable. Nous appelons le premier: radical et le second: désinence, en donnant à ces mots une valeur purement formelle, sans référence aucune ni à l'étymologie ni à la morphologie historique.

Pour analyser une telle forme, on doit disposer d'un lexique de radicaux des mots à flexion réguliere et d'un lexique de toutes les désinences de la langue. Dans ce dernier lexique, chaque désinence *-abat* sera accompagnée des codes 5 1 C 1 2 signifiant verbe de la première conjugaison, voix active, troisième personne du singulier, mode indicatif, temps imparfait. Par ailleurs, le lexique des radicaux comportera une forme *-am-*, accompagnée d'un lemme *amo* et d'un code 5 1, indiquant qu'il s'agit d'un verbe de la première conjugaison.

Dès lors, lorsque l'ordinateur est confronté au cours de la lecture du texte avec le mot *amabat*, il décompensera le mot en *am-abat*, il ira consulter son répertoire de désinences, y découvrira une finale *-abat* qui coincide avec celle qu'il est en train d'examiner. Cette finale est accompagnée des codes 5 1 C 1 2. Après quoi, l'ordinateur ira consulter son lexique de radicaux, y trouvera une forme *am-* accompagnée des codes 5 1 analogue d'une part à celle qu'il étudie et d'autre part analogue par les 2 premiers chiffres des codes à la finale *-abat* qu'il a lue dans son répertoire de désinences. Il peut donc conclure qu'il y a compatibilité entre les deux formes et il perforera une carte où se trouveront l'analyse complète et le lemme d'*amabat*.

Soit enfin la forme *praefert*. Son analyse est identique à celle du simple *fert* et son lemme peut être recomposé à partir de *fero* si l'on dispose d'un lexique de préverbes où l'on trouvera *prae*. Le cas est le même pour toutes les formes verbales composées.

Vous voyez donc qu'il est possible d'analyser toutes les formes latines à l'aide de quatre lexiques différents: un lexique de radicaux, un lexique de dé-

sinences, un lexique de formes et un lexique de préverbes.

Naturellement ce programme ne résout pas le problème des amphibologies qui foisonnent en latin: il donne toutes les analyses possibles des mots latins mais ne choisit pas. Ceci est un autre problème que nous aborderons plus tard, je pense.

Pour le moment, je cède la parole à M. Evrard qui va vous parler des lexiques de radicaux et de formes.

Exposé de M. E. Evrard.

La tâche qui m'incombe est la description du lexique des formes et du lexique des mots à flexion régulière. Ces deux lexiques remplissent ensemble un rôle analogue à celui d'un dictionnaire. Sans doute n'avons-nous pas jugé interessant de leur demander la traduction des formes d'un texte, ni même de leur lemme, mais ils sont conçus de manière à fournir, pour chaque forme, tous les lemmes et toutes les analyses lexicalement et morphologiquement possibles. L'accomplissement de cette tâche suppose d'ailleurs, selon des modalités qui seront décrites par M. Bodson, l'utilisation des deux autre lexiques (désinences et préverbes).

Observons dès à présent qu'un tel système décèle les homographies et, dans une certaine mesure, les polysémies, mais qu'il n'est pas capable de choisir, parmi les diverses possibilités, celle qui s'accorde avec le contexte.

Chaque lexique est formé d'une succession d'articles rangés dans un ordre alphabétique sur lequel nous aurons à revenir. En vue de leur utilisation par le programme d'analyse, les lexiques sont enregistrés sur deux piles de disques magnétiques placées respectivement sur le drive 1 (lexique des formes) et sur le drive 2 (lexique des mots à flexion régulière). Sur chacune de ces piles, les lexiques commencent au 2e cylindre. Chaque article occupe un secteur de 100 positions. Cette longueur fixe n'assure sans doute pas une densité maximum d'enregistrement, mais elle facilite grandement la constitution et la consultation du lexique. Au cours du chargement des lexiques sur les disques, l'ordinateur constitue automatiquement et enregistre sur le 1er cylindre du drive 1 une table des diagrammes initiaux, indiquant, pour chaque digramme, le secteur où il commence dans chacun des deux lexiques. Quant au 1er cylindre du drive 1, il est réservé au lexique des désinences. Cette disposition des enregistrements sur disque obéit à des motifs techniques: il s'agissait en effet de réduire le plus possible les mouvements des têtes de lec-

ture. Ainsi, l'exploitation du lexique des mots à flexion régulière suppose l'utilisation simultanée du lexique des désinences, ce qui est facilité par le fait qu'ils se trouvent sur deux *drives* différents.

*
* *

Le lexique des formes est constitué d'une série d'articles comportant chacun la forme considérée, son lemme et son analyse codifiée selon le système dont vous a parlé M. le Professeur Delatte (contenu des col. 54, 55, 56, 57, 58 et 61). Lorsque une même forme se prête à plusieurs lemmatisations et à plusieurs analyses, chaque possibilité distincte fait l'objet d'un article du lexique. Ainsi, pour *ac*, qui peut être une conjonction de coordination, ou une conjonction de subordination, notre lexique contient deux unités qui ont même forme et même lemme et qui ne diffèrent que par l'analyse. Notons toutefois que, pour distinguer ces deux *ac*, nous leur avons affecté un indice numérique. En vue de permettre la consultation automatique, les articles du lexique des formes sont rangés dans l'ordre alphabétique des formes. Vous trouverez un échantillon du lexique des formes dans le document intitulé *Echantillon du Lexique I.* Un bon exemple y est fourni par les deux articles *simul* qui, à la forme (2e col.), sont identiques, dont les lemmes (1re col.) se distinguent par les indices 1 et 2, mais auxquels sont ajoutées deux analyses distinctes: celle de conjonction de subordination (820) et celle d'adverbe (60000).

On pourrait concevoir un système où toutes les formes seraient analysées par le moyen d'un lexique constitué de la sorte.

Pour les mots invariables, il n'y a là aucun problème. Encore faut-il définir rigoureusement ce que l'on entend par mot invariable. Nos grammaires traditionnelles, par exemple, rangent tous les adverbes dans cette catégorie, oubliant que les adverbes de manière sont susceptibles de degrés de comparaison. Dans un lexique de formes, chaque adverbe de manière ferait donc l'objet de trois articles et occuperait trois secteurs.

Ceci nous conduit aux mots à flexion: noms, adjectifs, certains numéraux, pronoms, verbes. Ces mots occuperaient chacun, dans un lexique de formes, autant de secteurs qu'ils peuvent prendre de formes distinctes au cours de leur déclinaison ou de leur conjugaison. On devrait évidemment considérer comme distinctes les formes homographes mais ayant des analyses différentes: la forme *amici* du mot *amicus*, par exemple, fournirait la matière de trois articles, un pour chacune des trois analyses possibles (G. sg., N.pl., V.

pl.).

La constitution d'un tel lexique de formes serait, on s'en rend compte, une tâche considérable, puisque chaque mot variable devrait être préalablement ou décliné ou conjugué complètement. A un moment où nous pensions employer ce système, nous avions entamé les travaux préparatoires d'un programme qui aurait exécuté cette tâche automatiquement, du moins pour les flexions régulières.

Si nous avons abandonné cette voie, c'est que nous nous sommes rendu compte qu'un lexique de formes, s'il devait couvrir tous les types de mots, prendrait des dimensions excessives, compte tenu à la fois de la dimension modeste de notre ordinateur et des temps d'exploitation.

Pour le montrer, livrons-nous à quelques calculs fort simples. Chaque substantif compte douze formes. Chaque adjectif en compte 36 à chaque degré, ce qui conduit, si l'on additionne positif, comparatif et superlatif, à un total de 108 formes. Pour les verbes, ceux qui comportent un actif et un passif ont près de trois cents formes. Notons ici que, pour certaines formes, nous rencontrons des situations analogues à celles que nous signalions tout à l'heure à propos des adverbes de manière. L'infinitif futur, par exemple, compte douze formes, puisque le participe qui entre dans sa composition varie en genre et en cas et qu'il peut apparaître au nominatif ou à l'accusatif. Encore faudrait-il enregister ces formes de telle manière que l'ordinateur sache que l'auxiliaire peut y être sous-entendu. Quant aux verbes qui n'ont que l'actif et quant aux déponents et semi-déponents, le nombre de formes se situe aux environs de 150.

Par ailleurs, les relevés que nous avons faits jusqu'ici dans Sénèque, dans César et dans d'autres auteurs nous montrent que les noms représentent en général environ 35°/$_0$ du vocabulaire (c'est-à-dire des unités lexicales, et non des occurrences des mots dans les textes), les verbes comptant, eux, pour 30°/$_0$ environ et les adjectifs pour plus de 12°/$_0$.

Munis de ces indications, nous pouvons évaluer grossièrement l'étendue d'un lexique de formes couvrant un vocabulaire de 3000 mots. Le millier de substantifs qu'il comprendrait fournirait plus de 12.000 articles. Les 900 verbes en donneraient 250.000 et les 360 adjectifs, 36.000. En tenant compte des autres catégories de mots, nous arriverions ainsi à un lexique de 300.000 articles. Or, le vocabulaire de 3000 mots qui serait ainsi couvert, est notoirement insuffisant, comme nous le verrons dans la suite.

Ainsi donc, si le système fondé uniquement sur l'exploitation d'un seul lexique de formes parait séduisant à première vue et est théoriquement satis-

faisant, il se heurte, en pratique, à des difficultés excessives.

Tout d'abord, comme nous l'avons déjà fait remarquer, la préparation d'un tel lexique est une tâche fort lourde, où fourmillent les occasions d'erreurs. D'autre part, l'enrichissement d'un lexique d'un tel type est une tâche fort complexe, puisque l'adjonction d'un seul mot peut se traduire, en fait, par l'introduction de plusieurs dizaines, voire de plusieurs centaines d'articles, qui, dans l'ordre alphabétique global, ne se situent pas nécessairement en succession continue.

Enfin, et c'est là l'objection majeure, un lexique de plusieurs centaines de milliers d'articles comporte, eu égard aux caractéristiques de notre équipement, des difficultés techniques considérables. Son enregistrement exige des mémoires périphériques à très grande capacité (plusieurs dizaines de millions de positions). Ceci entraîne pour nous l'utilisation de plusieurs unités de mémoire (un lexique de 300.000 articles couvrirait 15 piles de disques): il en résulterait évidemment d'ennuyeux problèmes de manipulation.

De plus, il faut considérer le problème des temps d'accès et des conditions d'exploitation . Enregistré sur disques, notre lexique ne serait rentable que pour un très grand nombre de formes rangées en ordre alphabétique. En revanche, l'analyse d'un texte relativement court en suivant l'ordre du texte est exclue parce qu'elle absorberait un temps beaucoup trop considérable. Or, il nous semble que la lemmatisation et l'analyse ont tout avantage à se faire dans l'ordre du texte, puisque c'est le contexte qui permet au philologue d'opérer les choix nécessaires parmi les diverses possibilités que lui suggères la machine. Cet ensemble de considérations nous a amenés à rechercher les moyens de condenser au maximum les parties du lexique qui s'y prêtent et à limiter l'emploi du lexique de formes aux cas pour lesquels nous n'avons découvert aucune possibilité de condensation.

Ainsi réduit, le lexique des formes contient en premier lieu les mots invariables, pour lesquels la forme est identique au lemme: ce sont les prépositions, les conjonctions, de coordination et de subordination, les interjections et les adverbes, à l'exclusion des adverbes de manière régulièrement formés sur des adjectifs (type *beate, beatius beatissime* formé sur *beatus*,et type *grauius, grauissime* tiré de *grauis*). Nous avons vu tout à l'heure des exemples d'articles consacrés à ce type de mots.

En second lieu, le lexique des formes contient toutes les formes des mots variables qui n'obéissent pas à une flexion régulière. Pour ces formes, en effet, il est possible de faire une analyse en radical et désinence, mais une

telle analyse ne peut être ramenée à une règle générale: loin de simplifier le problème, elle apporte plutôt une complication supplémentaire. On trouvera donc, dans le lexique des formes, la plupart des adjectifs-pronoms ainsi que les noms, adjectifs et verbes irréguliers. Pour ces derniers toutefois, on n'a retenu en principe que les formes simples, les formes à préverbes étant lemmatisées et analysées par décomposition de la forme et recomposition du lemme, comme cela sera expliqué plus en détail ultérieurement.

L'échantillon du Lexique I que vous avez sous les yeux vous offre de bonnes illustrations de ce que je viens de vous exposer. Vous y trouverez, par exemple, à leur ordre alphabétique, plusieurs formes du verbe *esse: sim*, *sint, sis, sit,* etc. Chacune de ces formes se trouvent deux fois, parce qu'il nous a paru nécessaire de distinguer *sum 1* (verbe substantif) de *sum 2* (verbe auxiliaire). A chacun des articles, vous voyez, outre la forme et le lemme, une analyse en code qu'il vous est loisible de traduire en clair au moyen du tableau que vous a commenté M. Delatte.

*

* *

Pour les formes des mots à flexion régulière, il nous a paru économique de remplacer leur énumération par leur loi de formation: c'est là un processus que connaissent bien les hommes de science. Mais, ici comme ailleurs, il suppose la mise au point de tout un mécanisme délicat, dont le détail viendra mieux à sa place dans l'exposé sur le lexique des désinences, mais dont il faut bien que je dise dès maintenant quelques mots.

Si l'on veut lemmatiser et analyser une forme par les lois de la flexion régulière, on doit en premier lieu y pratiquer toutes les coupes possibles, puis, pour chacune de ces coupes, on doit vérifier si les deux parties obtenues peuvent être interprétées comme un radical et une désinence compatibles. Pour des raison d'ordre pratique, on commence cette opération par la fin, isolant successivement la finale d'une lettre, puis celle de deux lettres, etc.

Le lexique des désinences permet de distinguer celles qui, parmi ces finales, peuvent matériellement être des désinences. Or, chaque désinence, dans le lexique, est accompagnée d'une analyse. Retenons, dans cette analyse, la partie qui concerne la catégorie grammaticale et ses subdivisions, c'est-à-dire, dans notre code, le contenu des col. 54 et 55 (en négligeant toutefois, dans cette dernière colonne, le zoning qui indique soit les degrés de comparaison, soit les voix du verbe). Pour que la finale considérée soit une véritable désinence, il faut et il suffit:

1. que la partie qui la précède dans la forme à analyser soit un radical de mot;
2. que ce radical appartienne à la catégorie grammaticale par rapport à laquelle la finale peut avoir un rôle de désinence.

Eclairons ceci de deux exemples: l'analyse de *insulàs* suppose que l'on a reconnu *-as* comme une désinence de la 1re déclinaison (en code 11L00) et *insul-* comme le radical d'un mot de cette déclinaison. En revanche, le fait que *-es* est une désinence du subjonctif présent pour les verbes de la 1re conjugaison (en code 51B31) ne nous permet pas d'analyser *consules*, puisque *consul-* n'est pas un radical de verbe de la 1re conjugaison.

Les observations qui précèdent nous renseignent sur le contenu nécessaire et suffisant d'un lexique de mots à flexion régulière. Chacun de ses articles doit comporter:

1. un lemme, puisque c'est lui qu'on cherche;
2. un radical, que nous définissons très empiriquement comme la partie qui reste invariable au cours de la flexion;
3. les codes correspondant à la catégorie grammaticale et à sa subdivision.

Comme c'est habituellement le radical qui permet de choisir les articles du lexique utiles à l'analyse d'une forme, le lexique des mots à flexion doit être rangé dans l'ordre alphabétique des radicaux. L'échantillon du lexique II qui vous a été distribué vous donne des exemples concrets sur la disposition du lexique des mots à flexion régulière. La 1re unité concerne le lemme *lenis*, que vous voyez dans la 1re colonne. Le chiffre 5 est une indication de fréquence qui se trouve sur les cartes de chargement du lexique mais qui n'est pas enregistrée sur disque. Vient ensuite le radical *len* et, enfin, le code grammatical 24, qui signifie que *lenis* est un adjectif de la 2e classe en *-is*.

*

* *

Il faut noter que l'utilisation conjointe du lexique des mots à flexion régulière et du lexique des désinences se prête à toute une serie de jeux subtils. Tout d'abord, le lemme est représenté par une de ses formes, qui peut donc être reconnue immédiatement, sans passage par la décomposition en radical et désinence. D'autre part les grammairiens parlent à juste titre de désinences zéro: les cas de ce genre se reconnaîtront à ce que la forme avant toute soustraction de finale, est identique au radical.

Soit par exemple le verbe *lenire*, qui est enregistré en deuxième position sur la première page de l'échantillon: l'identité d'une forme à son lemme *lenio*, jointe au fait qu'il s'agit d'un verbe de la 4e conjugaison (code 54), permet de l'analyser comme une 1re pers. indic. pres. De même, l'identité de la forme au radical (*leni*), autorise une analyse d'impératif présent.

Par ailleurs, les numéraux variables empruntent presque tous leur flexion à l'une ou l'autre des classes d'adjectifs. Ce que l'on cherchera donc ici, ce ne sera pas une identité des catégories grammaticales relatives à la désinence et au radical, mais bien une convenance. Ainsi, à un radical caractérisé comme ordinal (32 en code) conviendront les désinences d'adj. de la 1re classe (21 en code). Il faut ajouter que dans ce cas, l'analyse que proposera l'ordinateur proviendra pour partie du lexique des désinences et, pour partie, du lexique des mots.

Quant aux adverbes de manière, l'ordinateur les analyse par les radicaux d'adjectifs, grâce à des désinences adverbiales (*-e, -etr, . . .*). Dans un cas de ce genre, l'ordinateur doit recomposer le lemme suivant des règles programmées.

Revenons par exemple au radical *len-*, qui nous est fourni par l'échantillon. A un certain moment, l'ordinateur, analysant *leniter*, décomposera cette forme en *len-iter*. Le tableau des désinences lui fournit, pour *-iter*, une analyse d'adverbe limitée aux cas où le radical appartient à un adj. de la 2e classe (codes 22 à 26). Comme c'est ici le cas, l'ordinateur recompose le lemme *leniter* et propose l'analyse 60000 (= adverbe).

Le cas des adjectifs de la 2e classe est, lui aussi, fort complexe: telle désinence convient dans tous les cas, telle autre ne se trouve que dans un certain nombre, très variable, de sous-catégories. Il a donc fallu imaginer tout un système de correspondance qui permette de passer en toute sûreté d'une désinence à un adjectif attribué à telle sous-catégorie. M. Bodson vous donnera sur ce sujet de plus grandes précisions.

Dans plusieurs cas, les distinctions nécessaires nous ont conduit à ajouter dans le lexique des mots à flexion régulière ce que nous avons appelé des notes additionnelles. Elles se trouvent dans des positions déterminées, qui se situent immédiatement après les codes des catégories grammaticales.

Une première série de notes additionnelles concernent les substantifs. Le code 3 caractérise les substantifs du masculin ou du féminin, tandis que le code 6 distingue les neutres. Corrélativement, le lexique des désinences caractérise par des codes distinctifs les désinences nominales qui sont liées au genre. Un exemple montrera le rôle de ces notes additionnelles. Soit la forme *nomines*. Si le lexique des mots se contente d'indiquer que *nomin* est le

radical d'un nom de la 3e déclinaison dont le lemme est *nomen*; si, par ailleurs, le lexique des désinences donne *-es* comme une désinence d'acc. pour les subst. de la 3e déclinaison, l'ordinateur, tout naturellement, analysera *nomines* comme un accusatif pluriel de *nomen*. C'est grâce aux notes additionnelles qu'il évite cette bévue: celles-ci lui font savoir que la désinence *-es* n'est utilisable que pour les mots qui ont le code 3 (c'est-à-dire les masculins et les féminins). Tout de même, le jeu des notes additionnelles écartera le danger d'analyser *iudica* comme un nominatif pluriel de *judex*.

Toujours en ce qui concerne les substantifs, il a fallu distinguer, par le code 7, les neutres en *-e, -al, -ar*, pour permettre à l'ordinateur, par exemple, d'éliminer l'analyse de *mare* comme abl. sg.

Retournons à notre échantillon du Lexique II. Au 5 article de la première page, vous voyez que, pour *lenitas*, rad. *lenitat-*, le code grammatical est 133, ce qui signifie qu'il s'agit d'un substantif (1) de la 3e déclinaison (3) du masculin ou du féminin, ce qui exclut toutes les désinences propres au neutre.

Le lexique contient aussi des notes additionnelles pour les articles consacrés aux verbes. Ceux-ci sont normalement représentés chacun par trois articles. Il a paru trop difficile, en effet, de formaliser en programme les relations entre les trois thèmes du présent, du parfait et du supin. Mais, pour éviter les fausses analyses et les fausses lemmatisations, il a fallu distinguer les radicaux du parfait par la note 4 et ceux du supin par la note 8, le radical du présent se caractérisant par l'absence de note. L'ordinateur vérifie donc, pour toutes les formes verbales, la convenance de l'analyse avec la note additionnelle du radical. Il évite ainsi de faire de *cepit* ou de *capto* un présent de *capio* (*-it* et *-o* sont en effet des désinences d'indic. présent, tandis que *cep-* et *capt-* sont des radicaux de verbes.)

A la première page de notre échantillon, vous lisez, par exemple, les trois unités du verbe *lenire* avec, respectivement, les radicaux et les analyses *leni*, 54 (verbe de la 4e conj.), *leni*, 544 (rad. du pf. d'un verbe de la 4e conj.) et *lenit*, 548 (rad. du supin d'un verbe de la 4e conj.). Un peu plus bas, vous rencontrez une unité qui a le radical *lenit-* et l'analyse 544.

Une autre note additionnelle signale certains types de substantifs. Un code particulier distingue en effet les *pluralia tantum*. Par ce moyen, l'ordinateur évite toute une série d'analyses et d'autre part, il oriente correctement l'analyse par le lemme. Illustrons ceci de deux exemples. La forme *armo* ne peut être anlysée ni comme datif ni comme ablatif de *armo*. Encore faut-il

que l'ordinateur sache que le neutre de la 2e décl, qui a pour radical *arm-* et pour lemme *arma* ne possède que les formes du pluriel. En revanche, l'ordinateur, devant analyser ***arma***, va détecter l'identité du lemme avec la forme, mais pour utiliser correctement cette observation, il doit savoir que les neutres *pluralia tantum* ont un lemme qui peut s'analyser comme $N_{,}$, V. ou Acc. pl.

C'est encore en raison des incidences propres à l'analyse par le lemme que nous avons attribué le code 2 aux verbes impersonnels. Pour les radicaux du présent, en effet, l'ordinateur interprète habituellement l'identité forme/lemme comme caractéristique d'une 1re personne de l'indic. présent. Comme les impersonnels ont pour lemme la forme de la 3e p.sg., on voit quelles erreurs pourraient se produire si l'attention de l'ordinateur n'était mise en éveil.

Telles sont les notes additionnelles qui contribuent au choix ou à l'élimination de certaines analyses.

D'autres notes additionnelles concernent le déroulement même du processus d'analyse et de lemmatisation.

On sait en effet que notre programme vise à donner non pas une analyse et un lemme, mais bien toutes les analyses et tout les lemmes possibles du point de vue du lexique et de la morphologie.

Il en résulte que, pour chaque forme, l'ordinateur parcourt la portion voulue de chacun des deux lexiques, et qu'il pousse la décomposition en radical et désinence aussi loin que le lui permet le lexique des désinences, suivant un système que M. Bodson vous expliquera.

Par ailleurs, comme cela a déjà été dit, pour réduire dans la mesure du possible l'étendue des lexiques, nous avons programmé l'analyse et la lemmatisation des formes verbales composées à partir des simples correspondants.

Dès lors, quand l'ordinateur a appliqué ses procédés d'analyse à une forme considérée comme simple, il examine si le groupe initial formé d'une, de deux, de trois etc. . . lettres s'identifie à un préverbe et permet de tester la valeur d'une décomposition de la forme en préverbe et forme simple. On notera d'ailleurs qu'une même forme peut se prêter à plusieurs décompositions (p.ex. *prae-terit* et *prater-it*). L'ordinateur reprend alors tout son cycle pour chacune des décompositions possibles. Il s'agit là d'un ensemble d'opérations fort long et qui, si l'on ne prend certaines précautions, peut en outre conduire à de grossières erreurs. Ici encore, un jeu de notes additionnelles nous a permis d'améliorer notre système.

Tout d'abord, pour beaucoup de mots invariables ou de formes de mots

à flexion irrégulière, nous pouvons affirmer que ni la décomposition en radical et désinence ni la décomposition en préverbe et forme simple ne peut donner de résultat. Tel est le cas pour *ad*, pour *adfer*, etc. Pour les formes de ce genre, un code 1 dans le lexique 1 entraîne une interruption du cycle et un passage au mot suivant. Notre échantillon ne nous en offre malheureusement pas d'exemple.

Il me reste enfin à parler de trois notes additionnelles communes aux deux lexiques et relatives à la décomposition des formes verbales composées. Ce sont les codes qui, dans les échantillons, sont représentés, à l'extrême droite, par un tiret, un zéro et un &.

L'ordinateur, qui procède par comparaisons et par identifications, peut traiter comme préverbes des débuts de mots qui matériellement, sont identiques à un préverbe, mais qui, pourtant, n'en jouent pas le rôle. Un cas frappant nous est fourni par *audes* qui, par analogie avec *aufert*, par exemple, est traité comme le subj. prés. d'un verbe *audare* composé de *au* et de *dare*. La solution nous est fournie par un code caractéristique des verbes qui ne peuvent jamais être utilisés que comme simples. Si nous en marquons *dare*, nous devons alors introduire dans nos lexiques les verbes tels que *circumdare* etc., composés où le simple garde la forme du simple intacte.

L'exemple de *dare* nous a amenés à réfléchir sur celui des composés de ce verbe qui sont formés en *-dere*, et sur tous les cas où le simple se modifie lorsqu'il entre en composition (*-ficere*, etc.). Par une raison d'économie, nous avons alors introduit dans nos lexiques des verbes simples fictifs, qui ne se trouvent qu'en composition et, pour empêcher à la fois les erreurs et les pertes de temps, nous les avons marqués d'un signe distinctif. Notre échantillon contient un exemple de ce type: *licio*, simple de *pellicio* etc.

Enfin, pour de très nombreux mots, on peut affirmer qu'aucune décomposition ne serait profitable. Une troisième note additionnelle a pour effet, pour les mots ainsi marqués, d'interrompre le processus avant la recherche des préverbes. C'est ce qu'indique, dans nos échantillons, le tiret, dont vous pouvez constater facilement la fréquence très élevée.

*

* *

Au terme de cette partie de mon exposé, je puis résumer en quelques mots les caractéristiques des lexiques qui fournissent les lemmes et contribuent à l'analyse:

1. Pour les mots invariables et pour les formes des mots variables irréguliers,

un lexique de mots donne la forme, le lemme et l'analyse.

2. Pour les mots à flexion régulière, un autre lexique donne le lemme, le radical et les caractéristiques grammaticales qui, avec l'aide du lexique des désinences, permettent de choisir le lemme et l'analyse corrects.
3. Dans l'un et l'autre lexiques ont été volontairement omises les formes des verbes composés qui peuvent s'analyser à partir du verbe simple. De même, on ne trouvera dans les lexiques aucun des adverbes de manière qui sont régulièrement formés sur des adjectifs. En revanche, des verbes simples fictifs permettent l'analyse de familles entières de composés.
4. Un jeu de notes additionnelles dirige l'ordinateur dans l'analyse et dans le choix du lemme, à la fois pour éviter certaines erreurs et pour raccourcir le cycle des investigations chaque fois que c'est possible sans dommage.

*

* *

Après avoir décrit la forme des lexiques, il me reste à parler de leur contenu, c'est-à-dire du vocabulaire que nous y avons introduit. Deux solutions étaient possibles: ou bien introduire un vocabulaire complet (et dans ce cas, nous aurions utilisé Forcellini, le meilleur des dictionnaires actuellement existants) ou bien opérer un choix.

La première solution, qui paraît la plus sûre, a pour nous un inconvénient majeur qui est le nombre de positions de mémoires occupées et le temps d'exploitation.

La seconde solution, à laquelle nous nous sommes arrêtés, comporte un risque dont nous sommes conscients. L'absence d'un article peut avoir pour conséquence, soit qu'une forme reste sans analyse, soit que la ou les analyses proposées par l'ordinateur soient toutes fausses.

Comme, de toute manière, le philologue doit revoir le travail de l'ordinateur pour éliminer les analyses possibles mais incompatibles avec le contexte, il nous est apparu que l'inconvénient n'avait aucune gravité.

Quant au choix lui-même, nous l'avons fait de manière fort empirique. Avant de mettre au point notre programme d'ordinateur, nous avions analysé manuellement plusieurs oeuvres de Sénèque: les trois Consolations et le De Clementia. Nous sommes partis du vocabulaire recueilli dans ces quelques oeuvres, puis, à mesure que nos travaux progressaient, nous avons enrichi nos lexiques en essayant de garder un juste équilibre entre l'excès, qui aurait alourdi le travail, et le défaut, qui aurait laissé trop de mots inexpliqués.

Au reste, le chargement des lexiques sur disques est une opération fort

rapide. Aussi nous est-il apparu qu'à côté du vocabulaire commun, nous pouvions introduire temporairement des mots particulièrement liés au sujet de l'oeuvre en cours d'analyse. C'est ainsi que, au moment d'analyser Cornelius Nepos, nous avons ajouté au lexique des mots toute une série de noms propres qui ont notablement accru l'efficacité de notre programme. Pour les noms communs et les autres mots, la chose paraîtra peut-être difficile. Toutefois, lorsqu'on traite une oeuvre d'une certaine longueur et d'une homogénéité suffisante dans les sujets traités, on peut commencer par analyser une partie et, à la lumière de cette expérience partielle, déterminer la liste des mots à ajouter pour la suite.

Quoi qu'il en soit, notre expérience nous a montré qu'avec des lexiques relativement restreints, on obtient des résultats très suffisants. On comprend aisément que la proportion des mots analysés varie avec le caractère des textes. Nous pouvons toutefois fournir des indications chiffrées que nous avons vérifiées sur différents types de textes.

Actuellement, nous utilisons un lexique de mots qui contient plus ou moins 8500 articles, c'est-à-dire 8500 mots. Quant au lexique de formes, il comporte un peu plus de 4600 articles qui représentent un millier de mots. Avec ce matériel, nous analysons environ 96% des mots d'un texte de prose.

Tel sont les caractères et le contenu des lexiques de formes et de mots. Mais pour que vous puissiez en suivre l'exploitation, il est temps que vous soient expliqués les deux autres lexiques.

Exposé de M.A. Bodson

L'essentiel de mon exposé sera consacré au lexique des désinences. Les informations qui figurent ont été préparées en fonction du programme d'analyse qui vous a été brièvement décrit. J'en rappelle une fois encore le principe. A un moment donné du cycle d'analyse, après l'exploitation du lexique de formes, nous demandons à l'ordinateur de découvrir, dans la forme à analyser, d'une part un radical possible, d'autre part, une désinence possible et de vérifier si ces deux éléments sont compatibles. Si le résultat de cette vérification est positif, l'ordinateur peut proposer une analyse et une lemmatisation de la forme.

I. Contenu de la table des désinences.

Vous aurez remarqué que, dans ce mode d'analyse, si le lexique des mots réguliers fournit le lemme, c'est à la table des désinences que l'on demande de fournir l'analyse morphologique complète. Cela signifie que toutes

les analyses morphologiques possibles doivent figurer à la table des désinences et cela explique que des désinences formellement identiques y sont reprises autant de fois qu'elles recouvrent une analyse différente: c'est ainsi que l'on trouve 32 désinences -A, 6 désinences -IA etc.

Notre première tâche en ce domaine a donc été d'opérer, en fonction du code utilisé au L.A.S.L.A., un relevé exhaustif de toutes les analyses admises pour les mots dits réguliers et de recouvrir chacune d'une désinence. Par exemple, l'analyse 12J appelle la désinence -A, 12L appelle aussi la désinence -A, etc. Nous avons relevé ainsi 934 désinences différentes sémantiquement, sinon formellement.

Cette opération n'a pas manqué, malgré son apparente simplicité, de poser certains problèmes.

A. En premier lieu, certaines analyses ne sont représentées par aucune désinence. Soit un verbe de la deuxième conjugaison, *moneo*. Son radical du présent, figurant au lexique des mots réguliers, est *mone-*; la deuxième personne du singulier de l'impératif présent actif est également *mone*. Il existe donc, même au point de vue particulier de l'ordinateur, des formes à désinence zéro. Il est évidemment impossible d'analyser une forme telle que *mone* selon le schéma qui vous a été décrit: la décomposition ne conduit à rien. Ces cas sont rares: outre celui que je viens de citer, il y a encore l'impératif présent de la quatrième conjugaison, *audi*, et l'ablatif singulier de la cinquième déclinaison, *die*.
Nous avons inclu les codes d'analyse de ces formes dans le programme lui-même, en demandant à l'ordinateur de les utiliser à deux conditions: a) que la forme à analyser soit identique à un radical du lexique des mots réguliers; b) que ce radical appartienne soit à un verbe de la deuxième conjugaison soit à un verbe de la quatrième conjugaison, soit à un substantif de la cinquième déclinaison.

B. Certaines analyses ne sont représentées par aucune désinence fixe: un bon exemple de cette situation est fourni par le nominatif singulier de la troisième déclinaison où, d'une part, le radical n'apparaît généralement pas et où, d'autre part, la désinence prend les formes les plus diverses. Il est donc impossible, ici encore, d'utiliser le schéma général de analyse. On aurait pu envisager d'inclure toutes ces formes insécables dans le lexique des formes irrégulières en les accompagnant de leur lemme et de leur analyse complète: *corpus* aurait fait l'objet de trois unités,

pater de deux unités. Cela aurait abouti à un gonflement considérable du lexique de formes

Pour parer à cet inconvénient, nous sommes partis d'une constatation: le fait qu'une forme, *rex* par exemple, soit identique au lemme *rex*, substantif masculin de la 3e déclinaison, autorise deux analyses de cette forme: nominatif singulier et vocatif singulier. Nous avons donc décidé d'opérer une comparaison de la forme à analyser, non plus avec les radicaux du lexique des mots réguliers, mais avec les lemmes des mots figurant dans les lexiques. L'ordinateur est programmé pour détecter les cas où l'identité forme-lemme autorise une ou plusieurs analyses. Ceci explique qu'un certain nombre de terminaisons soient absentes de la table des désinences. C'est ainsi que le -A du nominatif, du vocatif et de l'ablatif singulier de la première déclinaison ne figure pas dans la table. L'ordinateur sait, si j'ose dire, qu'une forme telle que *rosa*, identique au lemme *rosa*, 11 (substantif de la première déclinaison) peut accepter de ce fait trois analyses.

Un troisième problème s'est posé au moment de la réunion des analyses et des désinences: des analyses identiques sont représentées par des désinences différentes en fonction de certaines règles de la morphologie. Par exemple, l'analyse 13 L (substantif de la troisième déclinaison à l'accusatif pluriel) peut être représentée par les désinences -ES, -A ou -IA. Ce fait comporte de graves dangers de confusion, dangers que nous avons prévenus par l'usage des notes additionnelles dont vous a parlé M. Evrard. On retrouve donc dans la table des désinences le pendant des notes additionnelles qui figurent aux lexiques.

Je rappelle brièvement le sens de ces notes:

1: ne peut s'appliquer qu'à un adjectif de la 1ère classe;
3: ne peut s'appliquer qu'à un substantif masculin ou féminin de la 2ème ou la 3ème déclinaison;
6: ne peut s'appliquer qu'à un substantif neutre de la 2ème ou de la 3ème déclinaison;
7: ne peut s'appliquer qu'à un substantif neutre en -E, -AL, -AR de la 3ème déclinaison;
4: ne peut s'appliquer qu'à un thème du parfait;
8: ne peut s'appliquer qu'à un thème du supin;
11: ne peut s'appliquer qu'à un comparatif irrégulier.

Toutes ces notes servent à enrayer, à arrêter à temps, la logique im-

pitoyable de l'ordinateur.

D. Enfin, pour éviter un allongement excessif de la table des désinences, nous avons créé des désinences polyvalentes c'est-à-dire des désinences dont le programme précise qu'elles ont un champ d'application plus large qu'une seule catégorie. Je me limiterai ici à l'un ou l'autre exemple. M. Evrard vous a parlé des adjectifs de la 2ème classe. Il est bien évident qu'il eût été anormal de reprendre à propos de chaque sous-groupe, des désinences valables pour le groupe tout entier, le -IS du génitif singulier, par exemple. En pareil cas, une note d'analyse indique à l'ordinateur à quel sous-groupe la désinence convient. Ainsi, la désinence -VM sera réservée à certains sous-groupes, tandis que la désinence -IVM sera réservée à d'autres. Les désinences propres aux thèmes du parfait et du supin sont également valables pour toutes les conjugaisons. Un code est réservé à l'indication de cette propriété.

II. Ordonnance de la table.

La table de désinences occupe 187 secteurs de la mémoire à disque 1. Chacun de ces secteurs contient 5 désinences. Le premier nombre de trois chiffres figurant sur le listing donne le numéro du secteur où sont mémorisées les cinq désinences. Le chiffre suivant, de 0 à 4, indique la position précise de chaque désinence dans le secteur. Ainsi, la dernière désinence -A se trouve au secteur 004 et y occupe la 2ème position.

L'ordre dans lequel les désinences se suivent est le résultat de l'enchaînement de divers classements.

a) Le classement majeur a été opéré après alignement des désinences sur la droite. C'est le classement sur la lettre finale. On trouve d'abord toutes les désinences se terminant par -A, puis par -E etc. Ce classement est fondamental puisque l'ordinateur commence par isoler la dernière lettre du mot à analyser.

b) Dans ces groupes alphabétiques inverses, les désinences sont rangées en ordre de longueur croissante. Dans le groupe des désinences en -A, on trouve d'abord les désinences d'une lettre, puis de deux lettres, etc.
Ce classement est lui aussi imposé par le programme d'analyse, par lequel l'ordinateur, à la recherche d'une désinence, isole successivement 1, 2, 3, lettres à la fin du mot.

c) Au sein des groupes de longueurs identiques, les désinences semblables

forment des sous-groupes. Par exemple, dans les désinences en -A de deux lettres, on trouvera réunies en sous-groupes toutes les désinences -VA. C'est la fréquence relative de chacun de ces sous-groupes qui conditionne leur place.

d) Enfin, au sein du sous-groupe lui-même, les désinences sont rangées selon leur fréquence de manière que la première analyse fournie par la machine ait le plus de chance d'être la bonne. Par exemple, les désinences -VA sont reprises dans l'ordre accusatif pluriel, nominatif pluriel, vocatif pluriel.

III. Consultation de la table.

Lorsque l'ordinateur lit une forme, après sa recherche au lexique des formes irrégulières et invariables il entame parallèlement la consultation du lexique des mots réguliers et de la table des désinences.

Il commence par isoler le dernière lettre de la forme de manière à se trouver devant, d'une part, un thème possible, d'autre part, une désinence possible. Comme les désinences ne peuvent se terminer que par A, E, I, M, O, R, S, T et V, il s'interrompt si le mot se termine par une autre lettre.

Supposons qu'il ait isolé une de ces lettres. Le programme lui indique à quel endroit de la table il doit se reporter; l'adresse du début de chaque groupe alphabétique figure en effet en mémoire centrale; si le mot se termine par -A, la recherche commence au secteur 000 0, par -E, au secteur 020 2.

A partir de cet endroit, la table est normalement consultée en séquence; c'est la règle générale. Toutefois, cette consultation doit être contrôlée en trois points:

a) il faut indiquer à la machine quand elle doit isoler une lettre supplémentaire de la forme (ajustement).
b) il faut lui faire sauter les désinences qu'il est inutile d'examiner (saut).
c) il faut arrêter à temps la consultation (arrêt).

Ces instructions sont données à la machine par une série de signes conventionnels et d'adresses qui figurent à côté de certaines désinences et qui sont reprises dans le listing à la droite des désinences. Un exemple en montrera le jeu.

Supposons la forme *templa* dont l'ordinateur a isolé un radical TEMPL- et une désinence -A. L'ordinateur commence sa prospection au secteur 000 0, à la première désinence -A (12 L 00- note additionnelle 6) il trouve par ailleurs au lexique un mot dont le thème, l'analyse et la note additionelle

correspondent à ce qu'il a isolé. Il peut donc produire une analyse.

A ce moment, l'ordinateur, qu'il ait trouvé ou non une analyse, doit poursuivre sa recherche car on lui demande de fournir toutes les analyses possibles du mot. La première désinence -A n'étant accompagnée d'aucun lien de séquence, l'ordinateur suit la règle générale et consulte la table en séquence. Dans notre exemple, il examine donc toutes les désinences -A et trouve encore une analyse valable à la désinence 000 3 et la désinence 003 2.

Au terme des désinences -A, il va couper son mot en isolant cette fois deux lettres (-LA) et un nouveau radical (TEMP-). Cette opération est provoquée par la présence, sur la dernière désinence -A, d'un lien de séquence (004 2) qui signifie: passez à la désinence 004 2 mais en isolant une lettre supplémentaire.

Le groupe des désinences en -A de deux lettres comporte 6 désinences -IA, 1 désinence -DA, 1 désinence -RA, une désinence -MA, 3 désinences - VA. La recherche commence au groupe -IA, mais immédiatement se présente un lien de séquence d'un type nouveau, qui signifie: si la nouvelle désinence isolée n'est pas -IA, passez à la désinence qui figure en 005 3, mais sans isoler une lettre supplémentaire. Ceci évite à l'ordinateur de consulter en pure perte la tranche des désinences -IA. Cette restriction est commandée par le flag qui surmonte l'avant dernière position de l'adresse de renvoi.

Dans le cas de TEMP-LA, l'ordinateur arrive donc à la désinence 0005 3 (-DA). Cette désinence ne comporte pas d'analyse. Il s'agit en fait d'une désinence-relais, purement fictive, destinée à faire progresser la recherche vers la désinence -NDA (007 0). Ces relais sont nécessaires parce que l'ordinateur n'isole jamais qu'une lettre à la fois. Le lien de séquence qui figure à côté de -DA signifie: si la désinence isolée est -DA, passez à la désinence 007 0 et isolez une lettre supplémentaire. Les liens de séquence figurant à côté de -RA et -MA jouent le même rôle.

Dans notre exemple, l'ordinateur arrive donc à la désinence -VA, avec une coupe TEMP-LA. Le signe figurant à côté de la première désinence -VA lui dit: si la désinence isolée n'est pas celle-ci, arrêtez la recherche; si elle est celle-ci examinez-la puis continuez en séquence. En effet, à ce moment, si la désinence isolée n'est pas -VA, l'ordinateur n'a plus aucune chance de trouver dans le groupe des désinences de deux lettres une désinence, soit réelle, soit relais qui corresponde à ce qu'il a isolé (-LA).

A la dernière désinence -VA, un autre code lui indique que, s'il est arrivé jusque lá, il doit s'arrêter, car aucune désinence ultérieure ne se termine par(le groupe)-VA. Ce signe arrêterait par exemple la recherche dans le cas

d'une forme telle que CORNVA.

Un dernier type de renvoi peut encore intervenir (voir désinence 010 0). Ce lien de séquence accompagne certaines désinences isolées, servant de lien entre une désinence plus courte et une desinence plus longue. Il est donc nécessaire de renvoyer à la désinence plus longue. C'est ici le rôle de 0140. Mais par ailleurs, il est inutile d'opérer ce saut, si ce que l'on vient d'isoler ne correspond pas à ce qui figure dans la table. Si ce que l'on a isolé après l'examen de la désinence -IA n'est pas -TIA, il est inutile d'aller à -NTIA. C'est le rôle du flag qui figure sur la dernière position du renvoi.

LA TABLE DES PREVERBES

Toujours dans un but d'économie, nous avons évité de faire figurer dans les lexiques de mots et de formes les verbes composés, leur analyse et leur lemmatisation pouvant être opérés par programme. En effet, l'analyse d'une forme telle que *advenit* est exactement la même que celle de *venit;* quant à son lemme, il est analogue à celui du verbe simple précédé d'une préposition.

Au terme des divers processus d'analyse décrits jusqu'ici (analyse par le radical, analyse par le lemme, analyse par les deux lexiques), l'ordinateur se demande encore s'il ne se trouve pas devant un verbe composé. A cette fin, il isole successivement une, deux, trois, . . . lettres au début du mot et vérifie si la partie ainsi isolée ne représente pas un préverbe. Il doit pour cela disposer d'une nouvelle table, reprenant tous les préverbes de la langue.

Le travail de documentation a consisté à les rassembler, à les classer par longueur, en raison du processus d'investigation de la machine, et, au sein de chaque groupe, par ordre alphabétique, de manière à arrêter la recherche à temps.

Il a naturellement fallu tenir compte de divers phénomènes propres à la composition verbale:

a) assimilation et élision des préverbes (af - e-).
b) modification des thèmes verbaux en composition *(facere - ficere).* Ceci nous a amenés à inclure dans le lexique des verbes simples purement fictifs.
c) impossibilité, pour certains verbes simples, d'entrer en composition. En pareil cas, une note additionnelle signale cette impossibilité.

Revenons au travail de la machine et imaginons qu'elle ait isolé un préverbe possible en tête d'une forme quelconque. A ce moment, elle reprend à

propos de la forme simple isolée tout le processus d'analyse décrit auparavant. Si, d'une part, cette analyse l'a conduit à un resultat positif, si, d'autre part, l'analyse obtenue est une analyse de verbe, la machine peut fournir une analyse de la forme composée et, en quelque sorte, créer le lemme de cette forme en accouplant le lemme du simple et le préverbe isolé. Ce premier résultat n'interrompt pas la recherche, qui n'est interrompue qu'au moment où, en raison de la longueur du préverbe isolé, plus aucune découverte n'est possible.

Le programme d'analyse morphologique tel qu'il vous a été décrit est loin d'être parfait puisqu'il ne résout aucune amphibologie, ainsi qu'en témoignent les listings que vous avez sous les yeux.

Ceci est évidemment une grande faiblesse et provoque un ralentissement considérable du travail car le philologue est obligé de vérifier artisanalement toutes les analyses données par la machine. Conscients du retard apporté par la paradoxale nécessité de refaire le travail de l'ordinateur, nous nous sommes demandé s'il n'y avait pas moyen d'entrer plus profondément dans l'automatisation de l'analyse et d'obtenir par des critères purement formels que l'ordinateur lui-même, seul, résolve certains types d'ambiguités. Ainsi il nous parassait évident que le problème de *cum*, préposition ou conjonction devait pouvoir être réglé assez facilement. Nous pensions, en effet, que *cum* préposition est suivi d'un ablatif à un, deux, trois quatre mots après lui et qu'on ne pouvait trouver un *cum* conjonction suivi immédiatement d'un ablatif parce que, dans ce cas, l'amphibologie pouvait exister aussi pour un lecteur latin du moins à certaines époques. Mais une phrase du type suivant a réduit nos illusions à néant:

-in hoc omnes errore versamur ut non putemus ad mortem nisi senes inclinatosque jam vergere, CUM ILLO infantia statim et juventa omnisque aetas ferat (Cons. ad Marc., XXI, 7).

En réalité, tous les essais effectués montrent que le latin est une langue tellement dépourvue de structures fixes, tellement désarticulée (puisque l'ordre des mots, sans être indifférent, y est libre), qu'il nous a été impossible jusqu'à présent de découvrir des points de repère, des pivots, grâce auxquels on aurait pu appendre à l'ordinateur quelques règles qui lui auraient permis de reconnaître la fonction de chaque mot dans la phrase.

M. Evrard vous montrera tout à l'heure ces impossibilités formelles par quelques exemples.

En désespoir de cause, nous nous sommes demandé si nous ne pourrions pas nous limiter à tenter de distinguer les verbes des propositions subordon-

nées des verbes de propositions principales et de perforer sur la carte des verbes subordonnés, les codes des modes de subordination. Nous sommes partis de l'idée que seuls les verbes à un mode personnel ne comportent aucune amphibologie et qu'à ce titre, ils pourraient servir de points de repère dans la phrase. Nous avons alors établi trois règles qui, dans notre esprit, pouvaient résoudre certains problèmes.

1) Le premier verbe, après un point, est un verbe de proposition principale s'il n'y a pas de conjonction intercalée entre le point et le verbe.
2) Le verbe qui suit immédiatement une conjonction est le verbe commandé par cette conjonction.
3) Le verbe qui suit un verbe commandé par une conjonction est un verbe de proposition principale:
   a) si l'on ne trouve entre lui et le verbe de la subordonnée ni conjonction, ni relatif, ni adverbe relatif, ni aucun autre mode de subordination;
   b) si devant le verbe de la subordonnée ne se trouvent pas deux conjonctions qui se suivent sans verbe intercalaire à un mode personnel.

Ces conditions nous paraissaient assez strictes pour espérer aboutir à certains résultats et ne souffrir aucune exception. Hélas, l'examen de quelques phrases d'une Lettre à Lucilius nous a convaincus que ces règles ne servaient à rien, à l'exception de la première, mais qui couvre si peu de cas que cela ne nous a pas paru valoir la peine de faire un programme.

M. Evrard vous montrera, par quelques exemples l'impossibilité d'aller très avant dans ce domaine.

Ubi. . . rem esse in angusto vidit neque ullum esse subsidium, quod submitti posset, scuto ab novissimis uni militi detracto, quod ipse eo sine scuto venerat, in primam aciem processit centurionibusque nominatim appellatis reliquos cohortatus milites signa inferre et manipulos laxare iussit, quo facilius gladiis uti possent.

Caesar necessariis rebus imperatis ad cohortandos milites, quam in partem fors obtulit, decucurrit et ad legionem decimam devenit. Milites non longiore oratione cohortatus, quam uti suae pristinae virtutis memoriam retinerent neu perturbarentur animo hostiumque impetum fortiter sustinerent, quod non longins hostes aberant quam quo telum adigi posset, proelii committendi signum dedit.

Omnis in pilis Puteolorum turba consistit et ex ipso genere velorum Alexandrinas quamvis in magna turba navium intellegit: solis enim licet siparum

intendere, quod in alto omnes habent naves. Nulla enim res aeque adiuvat cursum quam summa pars veli; illinc maxime navis urguetur. Itaque quotiens ventus increbruit maiorque est quam expedit, antemna submittitur: minus habet virium flatus ex humili. Cum intravere Capreas et promunturium ex quo "alta procelloso speculatur vertice Pallas", ceterae velo iubentur esse contentae; siparum Alexandrinarum insigne indicium est.