Programme d'analyse morphologique et syntaxique du latin

La première tâche a été d'établir une liste de toutes les désinences possibles de la langue en ce qui concerne les déclinaisons et les conjugaisons. Chaque désinence est accompagnée d'un code traduisant sa valeur morphologique. Par exemple si l'on décide de codifier le substantif par 1, la première déclinaison par 1, et par ailleurs, le nominatif par 1, le vocatif par 2, une désinence en -a sera accompagnée des chiffres 111 et une autre désinence -a sera accompagnée des chiffres 112. Cet exemple montre que nous avons décidé de mettre en mémoire toutes les désinence homographes, chacune étant accompagnée d'un code indiquant l'analyse qui lève l'amphibologie de la désinence.

On pouvait à cet égard adopter une autre solution. Elle aurait consisté à n'introduire dans la mémoire que les désinencesqui ne prêtaient à aucune confusion. Nous n'avons pas retenu cette solution parce qu'il ne serait pas resté grand-chose en mémoire : pour la première déclinaison, trois formes seulement : -am, -arum, -as. Encore, deux de ces désinence peuvent-elles être confondues avec des formes de la conjugaison.

La solution que nous avons adoptée est parfaite dans la mesure où l'étude préliminaire a été complète,c'est-à-dire lorsqu' on n'a oublié de donner un code à aucune des désinences homographes. On apprendra avec surprise qu'en suivant cette méthode, on aboutit à la constitution d'une table de près de 1.000 désinences.

Le deuxième travail préliminaire a été de constituer deux lexiques. Le premier est composé de toutes les formes irrégulières des substantifs, des adjectifs, des adjectifs numéraux, des adjectifs-pronoms et des verbes; il comporte aussi une liste des adverbes de formation irrégulière, des prépositions, des conjonctions et des interjections.

Le deuxième lexique est le dictionnaire de base de l'ordinateur, celui auquel il se réfère pour découvrir à la fois la nature grammaticale du mot qu'il analyse et le lemme de ce mot, c'est-à-dire, la forme sous laquelle il apparaît au dictionnaire. Ainsi, si l'on codifie le substantif par 1, et la 1ère déclinaison par 1, rosa figurera dans le second lexique sous la forme rosa, ros- 11. Dès lors, lorsque l'ordinateur est confronté au cours de la lecture du texte avec le mot rosa, il décomposera le mot en ros-a, il trouvera dans son répertoire de désinence, à la forme -a, les codes 111 et 112, puis il ira voir dans son lexique et découvrira une forme ros- qui, ajoutée à la désinence qu'il a découverte, concorde avec la forme qu'il a lue. Cette forme est accompagnée dans le lexique des chiffres 11 signifiant qu'il s'agit d'un substantif de la 1ère déclinaison. L'ordinateur sortira donc deux cartes, l'une portant rosa 111, l'autre portant rosa 112.

Telle était du moins notre première intention. Mais il nous est apparu que la multiplication des cartes (dans certains cas, l'ordinateur sortirait 8 analyses différentes pour un seul mot) était un gaspillage et pourrait constituer un danger pour la suite du travail : n'oublie-

rait-on pas d'enlever toutes les cartes fautives d'un fichier ? Laisserait-on subsister la bonne analyse ?

Il nous est venu alors une autre idée. On sait que l'ordinateur dispose d'une machine à écrire grâce à laquelle il peut imprimer les résultats de ses propres recherches, avant de les perforer dans une carte. Cette caractéristique nous permet d'engager un véritable dialogue avec la machine. Le philologue se trouvera au pupitre de commande de l'ordinateur. Dès le moment où le travail commence, l'ordinateur transcrit sur sa machine à écrire l'analyse complète de chaque mot. Lorsqu'un mot est susceptible de plusieurs analyses, l'ordinateur, avant de transcrire la première analyse, écrit d'abord un astérique qui, conventionnellement, signifie : "Attention, plusieurs analyses du même mot vont se succéder sur le papier. "Ainsi l'attention du philologue est en alerte. Lorsque la première analyse a été écrite, l'ordinateur attend que le philologue appuie sur un bouton pour lui dire : "Cette analyse est la bonne; perforez-la et passez au mot suivant" ou qu'il appuie sur un bouton pour signifier : "Cette analyse est défectueuse, donnez l'analyse suivante." On me pardonnera cette personnification, cette humanisation de la machine. Mais c'est, je crois, la façon la plus commode de faire comprendre le travail de l'ordinateur.

Dans le cas où aucune analyse n'est possible pour une raison ou pour une autre, l'ordinateur attendra que le philologue tape lui-même l'analyse correcte. Ce dernier cas se présentera en particulier lorsque l'ordinateur, en consultant ses mémoires, ne trouvera pas le mot en question dans son dictionnaire.

Ceci nous ramène au problème du lexique dont l'ampleur est conditionnée par la place disponible sur les disques magnétiques. En fait, après la mise en mémoires du programme d'analyse, des désinences et du premier lexique, il reste encore la place nécessaire pour mémoriser 38.000 mots différents !

La tentation etait grande de faire une compilation du dictionnaire de Forcellini. Nous ne nous sommes pas arrêtés à cette solution pour plusieurs raisons. La première est d'ordre pratique : l'an prochain, paraîtra une reproduction anastatique de la dernière édition de ce dictionnaire; nous ne la possédons pas et il est à craindre que l'orthographe de beaucoup de lemmes n'ait été modifiée et qu'on ne distingue de nouveaux lemmes.

Par ailleurs, c'est un fait d'expérience que le lexique d'un auteur est assez limité. Pour autant que nous puissions en juger par nos propres travaux, on peut dire d'une manière générale qu'une œuvre de 8.000 mots ne comporte que 2.000 mots différents. Sans doute le vocabulaire est-il en rapport étroit avec le sujet traité : Sénèque n'emploiera pas les mêmes mots pour consoler Marcia et pour tenter de convaincre Néron de l'excellence de la Clémence. Il reste cependant un fonds commun important. Ceci a été mis en évidence pour toutes les langues et justifie la création de vocabulaires de base.

Pour le latin, Diederich publie une courbe de

compréhension des textes en rapport avec le nombre de mots différents : les 500 premiers mots couvrent 65 % de tous les textes latins, les 1.000 premiers mots, 76 %, les 1.500 premiers mots, 85 %, les 2.400 premiers mots, 90 %; les 3.500 premiers mots, 95 %, puis la courbe s'aplanit de plus en plus et le gain de compréhension est très maigre par rapport à l'extension considérable que doit prendre le vocabulaire.

Pour ces diverses raisons, il était inutile de reproduire tout le Forcellini. Nous nous sommes arrêtés alors à une solution intermédiaire. Puisque nous disposons déjà du vocabulaire des trois Consolations et du De Clementia, un tri général a permis de regrouper tous les mots différents de ces quatre œuvres. C'est ce lexique qui sera mis en mémoire et servira de base de départ pour nos travaux sur les Lettres à Lucilius. Ce lexique vient d'être achevé : il comporte seulement 3.850 mots différents pour plus de 30.000 occurences !

Cependant, pour profiter au maximum des possibilités de l'ordinateur, nous avons décidé de lui apprendre à enrichir son vocabulaire. Chaque fois qu'il rencontrera un mot nouveau qui lui sera dicté par le philologue, l'ordinateur l'enregistrera dans sa mémoire, à la place voulue et il s'en servira, dès que ce mot réapparaîtra dans le texte à analyser.

Tel est le programme qui vient d'être achevé. Il n'est pas sans intérêt de noter qu'il est d'une extraordinaire complexité

puisqu'il comporte plus de 1.000 ordres différents dont un grand nombre ont pour mission de permettre à l'ordinateur de choisir le circuit logique approprié à la forme qu'il doit analyser.

A cela s'ajoute que, parallèlement à l'analyse de chaque mot, l'ordinateur en assure automatiquement la référenciation complète (n° de chapitre, de paragraphe ou de vers, n° d'ordre du mot dans le paragraphe ou dans le vers, n° d'ordre du mot dans la phrase et enfin n° d'ordre du mot dans l'œuvre). Il utilise pour cela les codes de ponctuation qui figurent dans les cartes d'entrée où sont perforées les formes telles qu'elles se présentent dans le texte. Pour des raisons d'économie, ces formes sont perforées à raison de trois par carte. Par contre, il est indispensable pour la suite de la recherche que l'ordinateur constitue une carte par forme analysée et référenciée.

Tel est, sommairement décrit, le programme réalisé par l'équipe du Laboratoire. Nous en donnerons, l'an prochain, une description détaillée. Il est, d'ores et déjà, à la disposition des membres de l'Organisation.

L. Delatte.