Recherches sur le traitement automatique de la langue latine

Extrait de la Revue (R.E.L.O.)
IX, 1 à 4, 1973. C.I.P.L. - Université de Liège - Tous droits réservés.

INTRODUCTION

Les étapes du traitement automatique des données linguistiques

L'étude philologique des langues par les méthodes de l'informatique connaît depuis quelques années un développement considérable : un peu partout, les chercheurs ont recours à l'ordinateur pour sélectionner, classer et compter les faits de langue sur lesquels portent leurs investigations.

Il est évident que la compréhension d'une oeuvre, d'un auteur ou d'une langue nécessite une description systématique et des relevés complets des phénomènes morphologiques, syntaxiques et stylistiques contenus dans cette oeuvre, dans cet auteur ou dans cette langue.

Dans cette perspective, les *indices verborum* et les concordances apparaissent comme des instruments indispensables puisqu'ils fournissent des listes exhaustives du vocabulaire contenu dans une oeuvre. Cependant, la préparation d'un *index* par des procédés manuels demande de tels efforts que, le plus souvent, les chercheurs ont préféré se fonder sur une documentation partielle plutôt que de s'astreindre à des tâches fastidieuses.

Pour faire bien comprendre la difficulté de la réalisation manuelle d'un *index* ou d'une concordance, je rappellerai ce que disait, en 1951, le Père Roberto BUSA qui fut le premier, en Europe, à utiliser l'ordinateur pour établir des *indices* et des concordances :

" en se reportant à la nature des *indices verborum* et des concordances, je crois qu'on se rendra
" facilement compte de l'importance et de la difficulté du travail exigé pour leur réalisation.
" Découper le texte d'un auteur en phrases, le transcrire sur des cartes, retranscrire chaque carte
" autant de fois qu'il y a de mots dans la phrase, classer les mots en ordre alphabétique, de telle
" sorte que l'on ait dans un fichier et dans l'ordre adéquat, toutes, je le répète, matériellement toutes
" les prépositions, les conjonctions, les adverbes, les adjectifs, les substantifs, les verbes ... revient
" assurément à copier sept fois, vingt fois, et même quatre-vingts fois (c'est le cas du *Thesaurus*
" *Linguae Latinae*) l'oeuvre de la première à la dernière ligne avec en plus, pour chaque carte, addition
" de la citation et du mot de référence" (1).

Dans cette même perspective, B. QUEMADA, promoteur de la linguistique computationnelle pour la langue française, écrivait en 1959 à propos des concordances :

" l'énormité du travail matériel autant que le manque de vocation pour une tâche aussi fastidieuse
" semble avoir détourné les intéressés de la réalisation de tels travaux. La solution mécanisée est
" peut-être la seule à ce problème" (2).

Si le Père BUSA, lorsqu'il publia le premier *index* à l'aide des machines mécanographiques, insistait sur la difficulté et sur l'ampleur de la tâche manuelle, le Professeur DELATTE, lorsqu'il fonda l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur, mettait en évidence un autre inconvénient, plus grave, des *indices* réalisés manuellement :

" la plupart des erreurs qu'on découvre dans les *indices* faits à l'aide de fiches manuscrites sont iné-
" vitables et imputables à l'humaine nature incapable de copier correctement un texte (les manuscrits
" sont là pour nous en convaincre) incapable aussi de compter sans se tromper" (3).

C'est ainsi que, d'une part, la nécessité d'établir des *indices* et, d'autre part, la difficulté de le faire manuellement, ont amené la création de plusieurs centres qui se proposaient d'enregistrer et de traiter en ordinateur diverses langues et littératures.

D'une façon générale, les procédés mis au point peuvent se résumer en trois étapes :

1) *Enregistrement des textes sur un support compréhensible par les machines*

Cette première étape du traitement automatique des langues ne pose guère de problèmes : le texte est simplement perforé sur cartes mécanographiques ou sur bandes perforées; il peut également être transcrit sur des bandes magnétiques à l'aide d'une encodeuse. En fait, il s'agit d'une opération purement matérielle de reproduction du texte.

Cependant, en vue de l'indexation, l'ordinateur doit recevoir des précisions relatives à l'unité d'enregistrement puisque c'est à partir de celle-ci qu'il devra rechercher et identifier l'élément qui déterminera le classement dans l'*index*. La solution la plus fréquemment adoptée est de choisir le mot comme unité d'enregistrement. Toutefois, l'absence d'une définition satisfaisante du mot conduit souvent à des solutions empiriques. Ainsi, au Laboratoire d'Analyse statistique des Langues anciennes, on considère qu'en latin le mot peut être défini comme la plus petite unité située entre deux blancs typographiques.

Il y a cependant des exceptions tant au point de vue morphologique qu'au point de vue sémantique. Dans le premier cas, les formes composées de la conjugaison (AMATUS EST) sont considérées comme un seul mot, dans le second, des locutions telles que IUS IURANDUM ou RES PUBLICA sont traitées comme des expressions figées. Il en va de même des locutions prépositionnelles, adverbiales ou conjonctionnelles : QUAM OB REM, NON SOLUM SED ETIAM..., PRIUS ... QUAM.

Il est assez difficile d'établir au départ du traitement automatique d'une langue une liste exhaustive de ces expressions. A cet égard, deux principes énoncés par A.J. GREIMAS sont des auxiliaires précieux (4), même dans le domaine latin :

" a) l'élément sémiologique, dont les dimensions sont plus grandes que celles du thème peut être
" éprouvé par le test de *commutation* : il peut être remplacé dans le contexte par un autre élément
" ayant les dimensions d'un thème, sans que le sens général de la lexie en souffre (pomme de
" terre – carotte – navet);
" b) l'élément sémiologique, plus que le thème, possède, comme le mot, un ordre syntagmatique obli-
" gatoire. C'est un segment syntagmatique figé ".

Analyse du texte

Ici les solutions adoptées par les différents centres sont très diverses. Certains se contentent de ranger les formes du texte en ordre alphabétique. D'autres rattachent ces formes à la forme de référence qui se trouve dans un dictionnaire. D'autres encore font suivre chaque forme d'un embryon d'analyse morphologique. D'autres enfin, donnent à tous les mots du texte une analyse morphologique et syntaxique complète. Il est clair que, plus on enregistre d'informations à propos d'un mot, plus les possibilités d'exploitation des fichiers seront étendues.

Exploitation des données enregistrées

Il s'agit d'une phase purement automatique dont le but consiste à établir des listes de phénomènes dans un ordre adapté aux recherches que l'on doit entreprendre. Chaque type d'exploitation conduit

à des résultats imprimés (listings) qui répondent à un besoin spécifique. La liste que nous donnons ci-dessous est loin d'être exhaustive puisque les possibilités n'ont guère comme limites que les besoins du chercheur.

a) Classement du fichier en ordre alphabétique des lemmes et impression, de façon à obtenir un *index verborum* où toutes les formes d'un même mot seront réunies (ES, FUIT, SUNT seront regroupés sous le lemme SVM). Ce classement permet en outre de calculer la fréquence d'apparition de chaque lemme dans un texte donné.

b) Tris et comptages grammaticaux.

c) Concordances sélectives de tel ou tel phénomène linguistique.

Quel que soit l'aspect de la langue auquel on s'intéresse (sémantique, stylistique, psychologique), on peut extraire du fichier toutes les apparitions d'un phénomène avec un contexte plus ou moins long déterminé selon les besoins. Une application de ce procédé est la liste des verbes subordonnés rangés par modes et par temps, que l'on trouve dans les publications du Laboratoire d'Analyse statistique des Langues anciennes (5).

Cette phase d'exploitation se distingue des deux autres étapes à plusieurs points de vue.

Tout d'abord, elle n'est pas limitée dans le temps : un texte analysé se présente sous la forme d'un fichier enregistré sur cartes, sur bandes ou sur disques magnétiques de façon permanente. Il constitue dès lors une source d'informations, une banque de données que l'on peut consulter autant de fois que l'on veut et quand on veut.

En second lieu, l'exploitation dépend de la variété des indications qui sont codifiées dans le fichier. Plus la phase d'analyse est riche et complète, plus la phase d'exploitation offrira de champs d'investigations dans les domaines morphologique, syntaxique et stylistique.

Telles sont les trois opérations qui permettent entre autres de réaliser automatiquement des *indices.*

Assurément, la deuxième partie du travail, l'analyse, est la plus lourde car elle exige du philologue un travail long et fastidieux; elle impose, en effet, pour des raisons d'économie de place, de coder un grand nombre d'informations. C'est ainsi que dans le mémoire de licence que j'ai présenté naguère, j'ai analysé et codifié sans l'aide des machines les quelque 7500 mots de l'*Hercule furieux* de *Sénèque.*

Ce travail extrêmement lent s'est fait de la façon suivante :

1) analyse morphologique, syntaxique et stylistique de chaque mot du texte;
2) transcription sur les cartes de chaque analyse sous forme de codes;
3) perforation des informations portées sur les cartes;
4) vérification du fichier

Par le seul fait qu'elles imposaient plusieurs copies manuelles des informations, les trois premières opérations comportaient de nombreux risques d'erreurs dûs en particulier à la distraction et à la fatigue. La nécessité d'automatiser ce travail en le confiant à l'ordinateur apparaissait clairement.

En 1965, l'équipe du Laboratoire d'Analyse statistique des Langues anciennes réalisait un programme d'ordinateur pour l'analyse et la lemmatisation automatique du latin. Ce programme, qui a été utilisé jusqu'en 1970, fournissait pour chaque mot du texte traité tous les lemmes et toutes les analyses morphologiques possibles.

Les premiers résultats étaient certes fort encourageants puisque l'ordinateur traitait seul plus de 90 % des mots latins (6). Mais en proposant dans la plupart des cas plusieurs solutions, il laissait au philologue le choix définitif. Les recherches effectuées dans le but d'automatiser davantage l'analyse et la lemmatisation restaient sans résultat et, en 1968, le Professeur L. DELATTE écrivait :

" tous les essais effectués montrent que le latin est une langue tellement dépourvue de structure fixe,
" tellement désarticulée..., qu'il nous a été impossible jusqu'à présent de découvrir des points de
" repère, des pivots grâce auxquels on aurait pu apprendre à l'ordinateur quelques règles qui lui
" auraient permis de reconnaître la fonction de chaque mot dans la phrase" (7).

Du reste, la situation dans les autres langues n'était guère plus favorable. Les chercheurs les plus avancés dans le domaine de la traduction automatique devaient bien reconnaître que "le cerveau humain, bénéficiant de tout l'apport du réel, est mieux armé pour lever les ambiguïtés sémantiques et syntaxiques que ne le sera probablement jamais une machine (8)".

De nombreux spécialistes étaient portés à croire que l'étude des langues en ordinateur avait atteint des limites infranchissables; c'est ainsi que G. GOUGENHEIM écrivait :

" nous arrivons donc à cette constatation pénible que ce qui est important pour l'intercompréhen-
" sion, ... , pourra être exécuté par la machine, ... , tandis que ce qui est capital pour l'intercom-
" préhension exigera, si la langue de départ ne fournit pas d'indication, l'intervention d'un
" exécutant humain, qui devra se décider pour des raisons subjectives... à moins que l'on arrive à
" déterminer un nombre plus considérable de critères formels (9)".

Cependant, au cours de ces dernières années, la technologie des ensembles électroniques a évolué à grands pas; on a vu apparaître des ordinateurs beaucoup plus puissants dotés de nouvelles méthodes de stockage des fichiers. Les langages de programmation se sont considérablement développés et correspondent beaucoup mieux aux besoins spécifiques du traitement des faits de langue. Ce nouvel apport technique a conduit le Laboratoire d'Analyse statistique des Langues anciennes à revoir les procédés d'analyse automatique et à compléter ses programmes.

C'est pourquoi le travail que j'ai réalisé pour l'analyse automatique du latin a porté non seulement sur le programme, mais aussi sur l'organisation des lexiques qu'utilise l'ordinateur et, comme j'aurai l'occasion de le montrer, il m'a conduit à formaliser davantage le processus d'analyse. Avant d'exposer les méthodes que j'ai utilisées pour la lemmatisation et l'analyse automatique du latin, je rappellerai brièvement les notions de base du fonctionnement des ordinateurs et je définirai quelques termes techniques qui apparaîtront fréquemment dans cet exposé.

CHAPITRE I

Aspects techniques du traitement des données linguistiques

L'ordinateur se compose de plusieurs machines interconnectées qui ont chacune leur fonction propre dans le traitement des informations.

Chaque machine correspond à une des trois phases qui constituent le traitement de l'information.

1) La perception d'un enregistrement, d'un élément donné *(data)* se fait par l'intermédiaire d'unités de lecture appelées aussi unités d'entrée *(input)*.
Ces unités permettent de communiquer des données à l'ordinateur : ce sont les lecteurs de cartes et de bandes perforées et les armoires à bandes et à disques magnétiques.

2) L'analyse et l'identification d'une donnée s'accomplissent au niveau de l'unité centrale. Cette unité, dont le rôle est prépondérant, peut exécuter un certain nombre d'opérations arithmétiques et logiques. J'en parlerai bientôt.

3) La communication des résultats se fait par l'intermédiaire des unités d'écriture appelées aussi unités de sortie *(output)*. Ces unités sont l'imprimante, le perforateur de cartes, les bandes et les disques magnétiques.

Les armoires à bandes et à disques magnétiques qui peuvent fonctionner comme unités d'entrée et comme unités de sortie sont, en réalité, des mémoires périphériques, des mémoires auxiliaires, capables d'enregistrer et de stocker des millions d'informations. C'est grâce à leur capacité qu'il est possible d'enregistrer de façon permanente des lexiques et des fichiers que l'ordinateur utilise pour l'analyse des données linguistiques.

On comprendra mieux leur importance lorsqu'on saura que l'unité centrale, quelle que soit sa puissance, ne peut exécuter qu'un petit nombre d'opérations élémentaires et qu'elle ne peut fournir de réponse à une question que si elle a reçu au préalable tous les éléments qui lui permettent d'élaborer cette réponse.

Les opérations que l'ordinateur peut exécuter se répartissent en quatre groupes.

a) Les opérations d'entrée et de sortie destinées à la lecture et l'écriture des informations. Ces fonctions sont soit antérieures soit postérieures au traitement proprement dit, puisqu'elles ont comme rôle soit d'amener une donnée dans l'unité de traitement, soit de communiquer le résultat d'une série d'opérations.

b) Les opérations arithmétiques traditionnelles : addition, soustraction, multiplication et division.

c) Les opérations logiques permettant de comparer deux éléments et de décider en fonction du résultat de la comparaison s'il faut accomplir telle série d'opérations plutôt que telle autre.

Il y a entre les deux termes d'une comparaison six relations possibles.

Supposons les termes A et B.

A peut être égal à B (A = B)
différent de B (A¬ = B)
supérieur à B (A > B)
supérieur ou égal à B (A > = B)
inférieur à B (A < B)
inférieur ou égal à B (A < = B)

Ainsi, selon le rapport qui existe entre les deux termes, l'ordinateur peut choisir entre deux ou plusieurs séries d'opérations, celles qui conviennent plus particulièrement à la donnée qu'il traite.

d) Les opérations de transfert permettent de transporter dans une zone de la mémoire, la totalité ou une partie du contenu d'une autre zone. Ce type d'opérations est indispensable dans le traitement des données linguistiques. En effet, l'opération de transfert est la seule qui permet d'isoler facilement le mot ou la partie du mot que l'on doit plus spécialement examiner à un moment donné du programme.

Tous les langages de programmation possèdent d'une manière plus ou moins développée les instructions que nous venons d'énumérer. On pourrait donc penser qu'il n'y a pas lieu d'accorder trop d'importance au choix du langage. Ce serait une grave erreur : certains langages ont été conçus pour des applications purement mathématiques ou administratives et ils sont difficiles à utiliser dans le

traitement des faits de langue car les opérations logiques et surtout les opérations de transfert dont ces langages sont dotés, ne permettent pas d'examiner avec une souplesse suffisante le contenu d'une zone de mémoire.

Ces considérations techniques font apparaître deux critères importants qui guident le choix de l'ordinateur que l'on utilisera pour le traitement des données linguistiques.

Tout d'abord, cet ordinateur doit pouvoir être programmé dans un langage qui possède la souplesse nécessaire aux difficiles problèmes du traitement de l'information linguistique.

En second lieu, l'ordinateur devra être doté de mémoires périphériques à grande capacité - soit des disques, soit des bandes magnétiques - où l'on enregistrera les lexiques destinés à l'analyse et à la lemmatisation. Ce second critère est en étroite relation avec le fait que l'ordinateur doit recevoir tous les éléments qui lui permettent d'élaborer les réponses; sa caractéristique principale est d'enregistrer des millions de solutions et de reproduire à très grande vitesse une de ces solutions lorsqu'un problème connu se présente. La préparation des réponses apparaît dès lors comme la tâche principale dans la mise en place d'un système de traitement de l'information.

Etant donné l'importance des disques dans la suite de cet exposé, leur fonctionnement demande quelques mots d'explication.

Les disques magnétiques IBM 2311 que nous utilisons, contiennent deux mille pistes réparties en 200 cylindres concentriques. Chaque piste porte un numéro qui détermine le cylindre dont elle fait partie, ainsi que la position dans ce cylindre. La capacité maximum du disque est de 7.250.000 caractères.

Les disques magnétiques, destinés à l'enregistrement de fichiers volumineux, peuvent être utilisés de façon séquentielle, de façon aléatoire ou de façon indexée-séquentielle. Le mode séquentiel permet de lire (ou d'écrire) les enregistrements les uns à la suite des autres en commençant par le premier. Ce type d'utilisation convient bien lorsqu'il faut passer en revue tout le contenu d'un disque en traitant les informations dans l'ordre de leur succession naturelle.

L'utilisation aléatoire ou directe consiste à lire uniquement l'enregistrement dont on a besoin, quelle

que soit sa position sur le disque. On utilise à cette fin une *clé* qui permet de localiser l'information. La clé est composée, par exemple, du numéro du cylindre et du numéro de la piste où se trouve l'enregistrement, ainsi que du numéro d'ordre de cet enregistrement dans la piste.

Quant à l'utilisation indexée-séquentielle qui combine le mode séquentiel et le mode aléatoire, elle consiste à recourir à un *index* du fichier enregistré sur le disque, qui donne le numéro de la piste où se trouve chaque enregistrement. On peut comparer cet *index* à une table des matières extrêmement détaillée. Le repérage de l'enregistrement se fait grâce à une clé composée de chiffres aussi bien que de lettres et dont la longueur peut atteindre 256 positions. Cette clé doit être *unique* dans l'*index* aussi bien que dans le fichier, puisqu'elle ne doit jamais permettre d'accéder à plus d'un enregistrement. D'autre part, en indexé-séquentiel, la lecture d'un enregistrement n'est possible que si la clé que l'on a constituée *existe* dans l'*index*. Dans le cas où elle n'y est pas, l'ordinateur tente *sans y parvenir* d'exécuter l'instruction de lecture puis saute automatiquement une partie du programme. La nécessité d'avoir une clé *unique* et *existant* réellement dans le fichier, fait que toute lecture exécutée par l'ordinateur conduit obligatoirement à une *réponse exacte* sans aucun risque d'erreurs et sans qu'il soit besoin de procéder à des vérifications. On le voit, ce procédé offre simultanément des avantages de sécurité et de rapidité. Rapidité parce qu'il dispense du contrôle des réponses obtenues; sécurité parce qu'il n'y a aucune possibilité d'erreurs. Ces raisons expliquent que j'ai choisi cette solution pour stocker sur disques les lexiques destinés à l'analyse automatique.

Pour terminer ce paragraphe consacré au fonctionnement et aux possibilités des ordinateurs, je dirai quelques mots des moyens techniques que j'ai utilisés pour réaliser l'analyse du latin.

L'ordinateur sur lequel je travaille est un modèle 25 de la série IBM 360, doté d'une mémoire centrale de 24.576 positions, soit 24 k (1 k = 1024 positions). Il faut noter que 16.384 positions seulement sont utilisables puisque les programmes de gestion du système (programme superviseur) occupent 8.192 positions. Au point de vue des unités périphériques, notre ordinateur est équipé d'un lecteur-perforateur de cartes IBM 2540, d'une imprimante rapide IBM 1403 et de trois armoires à disques IBM 2311.

CHAPITRE II

Le système de traitement des textes latins

Le but du système est de faire produire par l'ordinateur la lemmatisation et l'analyse codée de toutes les formes d'un texte et, pour les formes ambiguës, tous les lemmes et toutes les analyses qui sont formellement possibles. Pour effectuer ces opérations, l'ordinateur fait subir à un texte cinq traitements successifs.

1. Dans une première étape, le texte entier, perforé sur cartes mécanographiques, est enregistré sur disques magnétiques. Chaque mot reçoit à ce moment un numéro d'ordre séquentiel destiné à permettre, au cours de la phase 4, le tri des formes dans l'ordre du texte. En outre, chaque forme est éventuellement suivie d'un code de ponctuation qui servira au calcul automatique de la référence. Si le texte analysé est une oeuvre dramatique, une troisième information peut également accompagner la forme : il s'agit d'un code qui permet de caractériser les différents rôles de la pièce.

   Les données peuvent se présenter de deux manières, soit à raison d'un mot par carte, soit à raison de trois mots par carte. Le dessin de ces cartes est le suivant :

   a) Un mot par carte :

Forme	colonnes 19-38
Code du rôle	colonne 64
Code de ponctuation	colonne 66

   b) Trois mots par carte :

	Forme	Ponctuation	Code du rôle
1ère zone	colonnes 1 -20	Colonne 21	
2ème zone	colonnes 22-41	Colonne 42	
3ème zone	colonnes 43-62	Colonne 63	
			Colonne 71

   Dans le mode d'enregistrement à trois mots par carte, on trouve en outre les deux particularités suivantes :

1) Le code symbolique du rôle est perforé une seule fois en colonne 71 et concerne tous les mots qui se trouvent sur la carte. Dès lors, chaque réplique doit obligatoirement commencer à une nouvelle carte.

2) Afin de faciliter la correction du fichier d'entrée, il a fallu prévoir un programme d'enregistrement qui accepte des cartes contenant moins de trois mots. Pour cette raison, l'ordinateur vérifie si chacune des zones de la carte contient une forme; si ce n'est pas le cas, la zone inoccupée est ignorée et l'ordinateur examine le contenu de la zone suivante ou éventuellement de la carte suivante.

La deuxième partie du travail consiste à classer les formes du texte en ordre alphabétique. Cette opération a pour but de diminuer la durée de la phase 3.

Au cours de la troisième phase, s'effectuent la lemmatisation et l'analyse. C'est à cet aspect, le plus complexe, que sera consacrée la majeure partie de cet exposé; pour l'instant, il suffit de savoir que les résultats de cette opération sont stockés sur disques magnétiques.

La quatrième étape du traitement consiste à reclasser les formes en ordre du texte à l'aide du numéro séquentiel qui leur a été attribué lors de la phase 1.

Durant cette dernière phase, chaque analyse est imprimée et fait l'objet d'une carte perforée. Les cartes que l'ordinateur perfore à ce moment se présentent selon le dessin suivant :

Col. 1	code carte
2 -17	lemme
18	éventuellement indice de lemme
19-38	forme
39-40	code d'oeuvre
41-43	n° du chapitre
44-47	n° du paragraphe
48-50	n° d'ordre dans le paragraphe
51-53	n° d'ordre dans la phrase
54-63	analyse

64	code symbolique du rôle (oeuvres dramatiques)
65	-
66	code de ponctuation
67-70	-
71-75	n° d'ordre dans l'oeuvre

Ce dessin de carte est celui que le L.A.S.L.A. utilise; il a fait l'objet d'un long article dans la *Revue* de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur (10).

En ce qui concerne les listings imprimés par l'ordinateur, on en trouvera des exemples dans l'annexe 3.

Au cours de la cinquième phase, chaque mot du texte reçoit une référence complète composée du numéro du chapitre, du numéro du paragraphe, du numéro d'ordre dans le paragraphe et dans la phrase et enfin du numéro d'ordre du mot dans l'oeuvre. Pour calculer cette référence, l'ordinateur se fonde sur les codes de ponctuation suivants :

2	fin de phrase
0	fin de paragraphe (ou fin de vers)
S	fin de paragraphe et fin de phrase
-	fin de chapitre
K	fin de chapitre et fin de phrase

Au début d'un travail d'impression et de perforation, il faut donner à l'ordinateur une carte d'initialisation qui lui indique le code d'oeuvre symbolique qui permet d'identifier le fichier ainsi que la référence de la première forme.

Cette carte d'initialisation présente le même dessin que les cartes-résultat perforées par l'ordinateur avec cette seule différence que les zones réservées au lemme, à la forme et à l'analyse sont inoccupées.

Les deux premières phases du système ainsi que les deux dernières ont un rôle secondaire puisqu'elles consistent, d'une part, à classer les mots d'un texte dans l'ordre le mieux adapté au traitement et, d'autre part, à présenter les résultats d'une manière telle que le philologue puisse facilement les examiner.

En revanche, la troisième phase occupe une place prépondérante puisque c'est à ce moment que chaque forme est lemmatisée et analysée. Pour effectuer ces opérations, l'ordinateur doit, au préalable, avoir reçu toutes les réponses aux questions qui lui sont posées ou, au moins, tous les éléments qui lui permettent d'agencer les réponses; c'est pourquoi il est nécessaire d'enregistrer dans les mémoires de l'ordinateur des lexiques, c'est-à-dire des listes de mots et d'analyses que le programme consulte constamment en vue d'identifier les formes d'un texte. En plus des données linguistiques, les lexiques peuvent contenir des informations d'ordre technique qui guident le travail de l'ordinateur.

Dès lors, le contenu des lexiques et leur présentation ont une importance capitale tant au point de vue de la valeur du système qu'au point de vue de la conception du programme.

L'étroite relation qui unit le programme et les lexiques fait qu'il est difficile, en ce qui concerne leur description, de donner la priorité à l'un plutôt qu'à l'autre. J'ai choisi de commencer par la description des lexiques parce qu'il me semble que cette façon de faire permettra de mieux comprendre le fonctionnement du programme.

CHAPITRE III

Le lexique

Le lexique doit contenir les éléments qui permettent à l'ordinateur de reconnaître un mot quelconque de la langue, de le rattacher à la forme type (lemme) qui figure au dictionnaire choisi comme ouvrage de référence et enfin de déterminer pour ce mot la ou les analyses morphologiques possibles, étant entendu que chaque analyse se compose de la catégorie grammaticale du mot, de sa sous-catégorie (déclinaison, conjugaison, etc.), du cas, du nombre, du degré de comparaison et du genre pour les formes à flexion nominale, de la personne, du nombre, de la voix, du mode et du temps pour les formes conjuguées.

On peut concevoir deux types de lexique radicalement différents.

1) Dans le premier type, chaque unité lexicale se compose de l'analyse et du lemme d'une forme et est destinée uniquement au traitement de cette forme. Dès lors, chaque mot de la langue doit être représenté, dans le lexique, par autant d'unités qu'il a de formes. Un tel procédé présente de graves inconvénients :

   tout d'abord, il impose au philologue de fournir, pour chaque mot, toutes les formes et toutes les analyses possibles. L'ampleur d'une telle tâche provoque des erreurs, des lacunes, des répétitions;

   par ailleurs, d'un point de vue technique, cette solution présente l'inconvénient d'exiger des mémoires périphériques gigantesques : elle impose en effet, d'enregistrer douze unités lexicales pour chaque substantif, trente-six pour un adjectif au positif et plus de deux cents pour les verbes de la conjugaison régulière. On peut estimer qu'un lexique du latin construit sur ce modèle doit contenir approximativement quatre cent mille unités pour permettre l'analyse de 90 % des textes.

2) Le second type de lexique se fonde sur la distinction entre la partie du mot qui reste constante à travers toute la flexion de ce mot et la lettre ou le groupe de lettres susceptibles d'affecter tous les lemmes appartenant à une même flexion. Nous désignons ces deux parties respectivement par les termes *radical* et *désinence* en donnant à ces mots une signification purement conventionnelle sans aucun lien avec la grammaire traditionnelle. Si l'on prend, par exemple, la forme AMABAT, on dira

que AM- est le radical de AMO, AMARE et -ABAT, la désinence de la troisième personne du singulier de l'indicatif imparfait actif des verbes de la première conjugaison.

Ce procédé de décomposition des formes conduit à la création de deux séries de données que l'ordinateur exploitera simultanément : un lexique des radicaux que nous appellerons dictionnaire de l'ordinateur et une table des désinences qui contient la morphologie de la langue.

Dans le dictionnaire, une unité lexicale est composée du radical d'un mot et de son lemme. Elle intervient dans l'analyse de toutes les formes de ce mot. Dès lors, avec un lexique relativement restreint (environ quatorze mille unités), l'ordinateur peut lemmatiser et analyser plus de 90 % des formes d'un texte latin.

Le principal intérêt du processus de décomposition réside dans l'économie appréciable d'unités lexicales qu'il permet de réaliser. Il existe toutefois plusieurs types de mots pour lesquels cette méthode est soit inutile, soit inapplicable.

Le premier type comprend les mots à flexion nominale dont la forme du nominatif ne peut être reconstituée à partir du radical : il en est ainsi, par exemple, de la plupart des substantifs de la troisième déclinaison (GENVS, GENER-, PATER, PATR). Pour ces mots, il a fallu créer, à côté de l'unité lexicale destinée à l'analyse des cas dérivés du radical normal, une unité permettant le traitement du nominatif, du vocatif et, pour les neutres, de l'accusatif.

Dans le deuxième groupe d'exceptions au processus de décomposition, on trouve les mots invariables qui sont représentés dans la langue par une forme unique et pour lesquels, par conséquent, une unité lexicale susceptible de fournir plusieurs analyses ne se justifie pas (NON).

Enfin, toutes les formes anomales sont traitées de la même manière que les mots invariables : chaque unité lexicale de forme irrégulière conduit à une seule analyse (EST).

La distinction entre formes décomposables et formes indécomposables, qui correspond à la distinction entre formes dérivées régulièrement d'un radical et formes anomales ou invariables, conduit à créer trois types d'unités lexicales.

1) Les unités lexicales destinées au traitement de toutes les formes d'un même mot; elles contiennent, outre le lemme, le radical et des codes destinés à la recherche de l'analyse dans la table des désinences.

2) Les unités qui conduisent à l'analyse des formes du nominatif, du vocatif et de l'accusatif neutre qui ne dérivent pas du radical normal. Ces unités sont susceptibles de produire plusieurs analyses (au maximum trois) sans utiliser la table des désinences.

3) Les unités destinées au traitement des mots invariables et anomaux. Elles ne peuvent produire qu'une seule analyse.

Les trois types d'unités lexicales ont la même configuration; leur plan général, matérialisé sur cartes perforées, est le suivant :

Col. 1	code caractéristique du type d'unité lexicale
Col. 2 -17	lemme
Col. 18	indice de lemme
Col. 19-38	radical ou forme
Col. 39-40	indice du radical
Col. 41-46	analyse codée, soit partielle, soit complète
Col. 47-49	notes additionnelles

Les paragraphes qui suivent sont consacrés à la description détaillée des unités lexicales. Toutefois, les notes additionnelles et l'indice du radical, en raison de leur caractère technique, sont provisoirement laissés de côté.

A. *L'unité-radical*

Ce type d'unité lexicale, le plus utilisé, est constitué des mots à flexion régulière et comprend normalement les substantifs, les adjectifs, les verbes, certains numéraux et quelques adjectifs-pronoms.

Le code caractéristique de ces unités est le chiffre 9 en colonne 1. Viennent ensuite le lemme et son indice* ainsi que le radical. Chaque radical est accompagné d'un début d'analyse grammaticale que nous appellerons sélecteur d'analyses parce qu'il est destiné à choisir les désinences qui peuvent affecter ce radical.

Ce sélecteur, utilisé pour la consultation de la table des désinences, occupe trois positions dans l'enregistrement (41-43). La colonne 41 indique la catégorie grammaticale du mot (substantif, adjectif, etc.) et la colonne 42 concerne la sous-catégorie grammaticale (déclinaison, classe d'adjectif, conjugaison). Ces deux colonnes contiennent obligatoirement une information codée, soit alphabétique, soit numérique. Quant à la colonne 43, elle est réservée à une indication complémentaire qui permet de distinguer certains groupes de mots à l'intérieur d'une même sous-catégorie grammaticale. Ainsi, pour les substantifs, le code de la colonne 43 se réfère au genre.

Pour certaines catégories grammaticales, le contenu de la colonne 43 peut être un blanc typographique. Ce blanc a autant de signification que les autres codes, il intervient comme eux dans le choix de la série des désinences. Nous le représenterons par le signe ƀ.

Voici, par exemple, des modèles d'unités lexicales pour les deux premières déclinaisons :

9	AMICITIA	AMICITI	11ƀ
9	DOMINVS	DOMIN	123
9	DONVM	DON	126

Le sélecteur 11ƀ de la première unité lexicale signifie que le radical AMICITI est un substantif de la première déclinaison et que seules les désinences caractérisées dans la table par le code 11ƀ peuvent intervenir dans l'analyse d'une forme de AMICITIA.

En ce qui concerne les substantifs de la deuxième déclinaison, le code 12ƀ ne peut suffire au choix des analyses. En effet, aux nominatifs et vocatifs singuliers et pluriels et à l'accusatif pluriel, les désinences des substantifs masculins ne sont pas compatibles avec les radicaux des neutres et, inversement,

---

* L'indice est un code numérique qui sert à distinguer les lemmes homographes. Ainsi, TEMPVS, le temps, reçoit l'indice 1, tandis que TEMPVS, la tempe, reçoit l'indice 2.

les désinences des substantifs neutres ne peuvent être associées à des radicaux de substantifs masculins.

L'indication du genre, en colonne 43, a pour rôle d'écarter de fausses analyses. En voici deux exemples : la forme DOMINA se décompose en un radical DOMIN- et une désinence -A. Or, DOMIN- est notamment le radical de DOMINVS et A est, entre autres choses, la désinence du nominatif, du vocatif et de l'accusatif pluriels des substantifs neutres. Dès lors, si aucun code ne distingue les masculins et les neutres, l'ordinateur, doté d'une logique purement formelle, va rattacher la forme DOMINA au lemme DOMINVS et fournir trois analyses erronées. De la même manière, l'ordinateur analysera la forme SERVVM comme le nominatif et le vocatif singuliers de SERVVS.

La nécessité d'éliminer de telles erreurs a entraîné la création de soixante-deux sélecteurs d'analyses et, parallèlement, de soixante-deux groupes de désinences. Au niveau des catégories grammaticales, il y a douze sélecteurs pour les substantifs, quatorze pour les adjectifs, trois pour les numéraux, deux pour les adjectifs-pronoms et trente et un pour les verbes.

Dans le tableau qui suit, chaque sélecteur est accompagné de la catégorie et de la sous-catégorie grammaticale à laquelle il s'applique ainsi que d'un lemme type et de son radical.

1. SUBSTANTIFS

Sélecteur	Catégorie	Lemme	Radical
11b	1ère déclinaison	AMICITIA	AMICITI
123	2ème déclinaison (masculins)	SERVVS	SERV
126	2ème déclinaison (neutres)	DONVM	DON
13b *	3ème déclinaison (subst. -O, ONIS)	MONITIO	MONITI
131 *	3ème déclinaison (subst. -OR)	GENITOR	GENIT
132 *	3ème déclinaison (subst. -AS, ATIS)	AVCTORITAS	AVCTORIT
133	3ème déclinaison (masculins et féminins)	REX	REG
136	3ème déclinaison (neutres)	ITER	ITINER
137	3ème déclinaison (subst. en -E, -AL, -AR)	MARE	MAR
143	4ème déclinaison (masculins et féminins)	MANVS	MAN
146	4ème déclinaison (neutres)	GENV	GEN
15b	5ème déclinaison	DIES	DIE

* Voir à ce sujet le premier cas particulier dans le chapitre consacré à la table des désinences (p. 43).

2. ADJECTIFS

Sélecteur	Catégorie	Lemme	Radical
2ǂƀ	comparatifs sans positif	DETERIOR	DETER
2- ƀ	superlatifs sans positif	DETERRIMVS	DETERRIM
21ƀ	1ère classe	BONVS	BON
22ƀ	2ème classe	VETVS	VET
23ƀ	2ème classe	ACER	ACR
24ƀ	2ème classe	FORTIS	FORT
25ƀ	2ème classe	FELIX	FELIC
26ƀ	2ème classe	MEMOR	MEMOR
2Fƀ	2ème classe (comparatifs irréguliers)	IVVENIS	IVN
2Jƀ	1ère classe (superlatifs irréguliers)	POSTREMVS	POSTREM
2Kƀ	2ème classe (superlatifs irréguliers)	PAVPERRIMVS	PAVPERRIM
2Lƀ	2ème classe (superlatifs irréguliers)	CELERRIMVS	CELERRIM
2Mƀ	2ème classe (superlatifs irréguliers)	DIFFICILLIMVS	DIFFICILLIM
2ψƀ	2ème classe (superlatifs irréguliers)	VBERRIMVS	VBERRIM

3. NUMERAUX

Sélecteur	Catégorie	Lemme	Radical
32ƀ	ordinaux	DECIMVS	DECIM
33ƀ	distributifs	BINVS	BIN
34ƀ	multiplicatifs	DVPLEX	DVPL

4. ADJECTIFS-PRONOMS

Sélecteur	Catégorie	Lemme	Radical
42ƀ	possessifs	NOSTER	NOSTR
48ƀ	indéfinis	NVLLVS	NVLL

5. VERBES

Sélecteur	Catégorie	Lemme	Radical
51b	1ère conjugaison (thème du présent)	AMO	AM
514	1ère conjugaison (thème du parfait)	AMO	AMAV
518	1ère conjugaison (thème du supin)	AMO	AMAT
52b	2ème conjugaison (thème du présent)	MONEO	MONE
524	2ème conjugaison (thème du parfait)	MONEO	MONV
528	2ème conjugaison (thème du supin)	MONEO	MONIT
53b	3ème conjugaison (thème du présent)	ACVO	ACV
534	3ème conjugaison (thème du parfait)	ACVO	ACV
538	3ème conjugaison (thème du supin)	ACVO	ACVT
54b	4ème conjugaison (thème du présent)	VESTIO	VESTI
544	4ème conjugaison (thème du parfait)	VESTIO	VESTI
548	4ème conjugaison (thème du supin)	VESTIO	VESTIT
55b	4ème conjugaison mixte (thème du présent)	CAPIO	CAP
554	4ème conjugaison mixte (thème du parfait)	CAPIO	CEP
558	4ème conjugaison mixte (thème du supin)	CAPIO	CAPT
564	anomaux (thème du parfait)	AVFERO	ABSTVL
568	anomaux (thème du supin)	AVFERO	ABLAT
5Jb	1ère conjugaison dépon. (thème du présent)	HORTOR	HORT
5J8	1ère conjugaison dépon. (thème du supin)	HORTOR	HORTAT
5Kb	2ème conjugaison dépon. (thème du présent)	VEREOR	VERE
5K8	2ème conjugaison dépon. (thème du supin)	VEREOR	VERIT
5Lb	3ème conjugaison dépon. (thème du présent)	LABOR	LAB
5L8	3ème conjugaison dépon. (thème du supin)	LABOR	LAPS
5Mb	4ème conjugaison dépon. (thème du présent)	MENTIOR	MENTI
5M8	4ème conjugaison dépon. (thème du supin)	MENTIOR	MENTIT
5Nb	4ème conjugaison mixte (thème du présent)	PATIOR	PAT
5N8	4ème conjugaison mixte (thème du supin)	PATIOR	PASS
5Sb	2ème conjugaison semi-dépon. (thème du présent)	AVDEO	AVDE
5S8	2ème conjugaison semi-dépon. (thème du supin)	AVDEO	AVS
5Tb	3ème conjugaison semi-dépon. (thème du présent)	FIDO	FID
5T8	3ème conjugaison semi-dépon. (thème du supin)	FIDO	FIS

Ce tableau appelle quelques observations.

1) Les adjectifs dont le superlatif ne se termine pas par le suffixe -ISSIMVS, sont représentés par deux unités lexicales; la première est destinée au traitement des formes du positif et du comparatif, la seconde sert à l'analyse des formes du superlatif. Ainsi, on trouve pour l'adjectif FACILIS les deux unités suivantes :

9	FACILIS	FACIL	24b
9	FACILIS	FACILLIM	2Mb

Le sélecteur d'analyse diffère d'une unité à l'autre. Les codes 24b indiquent que le radical FACIL- est destiné à l'analyse des formes du positif et du comparatif, tandis que les codes 2Mb conduisent au choix des désinences -VS, -A, -VM des superlatifs.

2) Les adjectifs de la première classe terminés par les suffixes -FICVS, -DICVS et -VOLVS font également l'objet de deux unités. Ainsi, pour l'adjectif MALEFICVS, le lexique contient :

9	MALEFICVS	MALEFIC	21b
9	MALEFICVS	MALEFICENT	21b

Le sélecteur d'analyse est le même dans les deux unités puisque les désinences du comparatif et du superlatif sont les désinences normales des adjectifs de la première classe (-IOR et -ISSIMVS).

3) Les cinq derniers sélecteurs d'adjectifs sont réservés aux superlatifs irréguliers.

4) Chaque verbe est normalement représenté par trois unités lexicales, car il est difficile voire même impossible de décrire en quelques règles les relations existant entre les thèmes du présent, du parfait et du supin. Pour cette raison, chaque thème est considéré comme un radical et fait l'objet d'une unité. C'est ce que montre l'exemple suivant :

9	AMO	AM	51b	thème du présent
9	AMO	AMAT	518	thème du supin
9	AMO	AMAV	514	thème du parfait

Dans ce cas, le code de la colonne 43 sert à déterminer les temps formés à partir des différents thèmes.

5) Les adverbes de manière ne sont pas traités comme les autres mots invariables. En effet, la plupart de ces adverbes sont formés régulièrement à l'aide des désinences -E ou -ITER ainsi que -IVS et -ISSIME à partir du radical d'un adjectif. Dès lors, puisque des désinences d'adverbes sont susceptibles d'affecter des radicaux d'adjectifs, c'est le radical de l'adjectif qui sert au traitement. Soit à analyser l'adverbe AEQUE, l'ordinateur trouve l'unité lexicale de l'adjectif :

9 AEQVVS AEQV 21b

Le sélecteur 21b conduit l'ordinateur à une désinence -E analysée 60000, c'est-à-dire adverbe de manière au positif. A ce moment, l'ordinateur transfère dans la zone du lemme le radical AEQV- et lui adjoint la désinence -E. Ainsi, le lemme est recréé automatiquement.

Ce système permet l'économie d'un grand nombre d'unités lexicales. En effet, si on avait traité les adverbes de manière comme les autres mots invariables, il aurait fallu trois unités par adverbe (positif, comparatif et superlatif).

Au terme de cette description, je résumerai en quatre points les particularités du premier type d'unités lexicales.

a) Chaque unité, caractérisée par le chiffre 9, se compose d'un lemme, d'un radical et d'un sélecteur de désinences.

b) Chaque unité doit conduire à l'analyse de toutes les formes régulières du mot qu'elle représente c'est le principal avantage du processus de décomposition : il permet d'analyser près de 500.000 formes différentes avec un lexique restreint.

c) Un même lemme peut être représenté par plusieurs unités lexicales lorsque la morphologie le permet ou l'impose.

d) Une unité lexicale peut intervenir dans la formation de lemmes absents du lexique : c'est ainsi que sont créés les adverbes de manière.

B. L'unité lexicale à générateur

L'unité-générateur intervient dans le cas où le nominatif singulier, le vocatif singulier et l'accusatif neutre singulier d'un substantif ou d'un adjectif ne peuvent être analysés par le procédé de décomposition en radical et désinence.

Cette unité s'applique aux catégories suivantes :

a) les substantifs de la deuxième déclinaison du type APER;

b) tous les substantifs de la troisième déclinaison, à l'exception de ceux en -O, -ONIS, -AS, -ATIS et -OR, -ORIS;

c) les adjectifs en -ER de la première classe et les adjectifs de la deuxième classe des types VETUS, ACER et AUDAX*.

Prenons l'exemple du mot AGMEN. Il est clair que les formes du nominatif, du vocatif et de l'accusatif ne peuvent être traitées à partir du radical AGMIN-.

Plusieurs solutions sont possibles pour résoudre cette difficulté.

La première consiste à isoler dans le mot un radical compatible avec toutes les formes fléchies. Ainsi, pour AGMEN, le radical AGM- peut intervenir dans l'analyse de toutes les formes, à condition de créer les désinences -EN, -INIS, -INI, -INE, etc. Cette solution présente un grave inconvénient : elle multiplie exagérément les séries de désinences puisqu'elle limite l'utilisation d'une série à des cas d'espèce.

A titre d'exemple, en se limitant aux principaux types de la troisième déclinaison, il faudrait créer des désinences pour les mots en -O, -INIS, -VS, -VRIS, -ANS, -ANTIS, -X, -GIS, -OR, -ORIS, -X, -CIS.

---

* Il est possible d'étendre davantage le processus d'analyse par générateur. C'est ainsi que sont traités, par exemple, les adjectifs-pronoms NOSTER et VESTER.

La deuxième solution consiste à enregistrer les deux ou trois analyses que l'ordinateur ne peut produire par le radical comme des unités-formes. Cette façon de procéder provoque évidemment un fort accroissement du lexique puisque chaque analyse possible doit faire l'objet d'une unité lexicale.

La solution que nous avons finalement adoptée élimine le recours à la table des désinences et permet d'éviter un gonflement excessif des lexiques. Elle consiste à enregistrer, en plus de l'unité-radical, une unité affectée d'un code spécial en colonne 45. Ce code indique à l'ordinateur le nombre d'analyses différentes qu'il peut produire à partir de l'unité (deux analyses pour les substantifs masculins et féminins, trois analyses pour les substantifs neutres). Nous avons donné à ce code le nom de générateur d'analyse.

Voici deux exemples d'unités lexicales à générateur :

9	AGMEN	AGMEN	13b	3
9	CAESAR	CAESAR	133	2

Ces exemples font apparaître la souplesse de l'unité à générateur :

1) ces unités ont le même code que l'unité-radical. Par conséquent, dans certains cas bien précis, elles pourront se substituer à cette unité;
2) par ailleurs, l'information que l'on trouve à la suite du lemme est la forme complète du nominatif. Dans le second exemple, cette forme est identique au radical. Dès lors, l'unité lexicale est ambivalente;
3) le sélecteur d'analyse ne joue, en principe, aucun rôle dans le choix des désinences puisque l'unité-générateur n'utilise pas la table des désinences. Cependant, pour les mots dont le radical est identique à la forme du nominatif, le sélecteur retrouve sa fonction normale. En effet, si l'ordinateur rencontre la forme CAESAR, il produit deux analyses (nominatif et vocatif singuliers) sans utiliser la table des désinences. Par contre, s'il rencontre une forme fléchie (CAESARIS, CAESARE ..), CAESAR- est identifié comme un radical; les notes 133 jouent normalement le rôle de sélecteur de désinences, tandis que le générateur est sans effet;
4) la valeur du générateur est 2 pour les substantifs masculins et féminins et 3 pour les substantifs neutres;

5) pour les formes d'adjectifs et d'adjectifs-pronoms, l'indication du cas ne suffit pas, il faut y ajouter le genre. A cet effet, l'ordinateur utilise, d'une part, la colonne 42 où est codifiée la classe des adjectifs et, d'autre part, quelques informations rangées en mémoire centrale où se trouvent les codes des genres.

En conclusion, l'unité à générateur évite la surcharge des lexiques et de la table des désinences; elle présente la particularité de faire appel à des éléments stockés dans le lexique et dans la mémoire centrale de l'ordinateur. En outre, elle est compatible avec l'unité-radical qu'elle peut, dans certains cas, remplacer.

L'unité-forme

Cette dernière catégorie d'unités lexicales relève du premier type de lexique dont j'ai parlé au début du chapitre : chaque unité représente une forme; elle ne peut produire qu'un seul lemme et qu'une seule analyse. Ce type d'unité est réservé au traitement des mots invariables et des mots anomaux. Son schéma est le suivant :

Col. 1	rien
Col. 2 -18	lemme
Col. 19-38	forme
Col. 41-46	analyse morphologique complète

Ainsi, pour la négation NON, l'unité lexicale est :

bNON	NON	68000

Sans doute, aurait-il été possible de généraliser le processus de décomposition en radical et désinence et de l'étendre aux mots invariables et irréguliers. Il n'y a, en effet, aucun empêchement à considérer une forme invariable comme un radical que l'on peut affecter d'une désinence zéro : on obtiendrait de la sorte un procédé d'analyse unique et un programme très simplifié.

Mais, la réduction de l'analyse à un processus unique impose, pour le traitement de chaque forme invariable ou anomale, une consultation inutile de la table des désinences. En effet, dans le cas des formes invariables ou anomales, aucune économie d'unités lexicales ne peut être réalisée et la forme se suffit à elle-même, tant pour l'analyse que pour la lemmatisation.

Au reste, si la décomposition apparaît grammaticalement impossible et logiquement injustifiée, elle risque, en outre, de conduire à de fausses analyses. C'est pourquoi, trois catégories de formes sont enregistrées dans le lexique avec leur analyse complète.

1. *Les invariables*

Grammaticalement, il n'est pas possible de distinguer un radical et une désinence pour les conjonctions, les prépositions, les interjections, les numéraux indéclinables et les adverbes autres que les adverbes de manière.

D'autre part, les mots invariables, par définition, ne sont jamais représentés que par une forme unique; par conséquent, une seule unité lexicale suffit à les analyser. Il est dès lors plus logique que l'analyse puisse être produite sans consulter la table des désinences.

Soit, par exemple, la forme AD dont l'unité se présente comme suit :

ƀAD	AD	70300

Lors du traitement de cette forme, le lexique apporte simultanément le lemme (AD) et l'analyse complète (70300 = préposition régissant l'accusatif) sans qu'il soit nécessaire de recourir à la table des désinences.

Quant aux mots invariables qui, par application des règles d'euphonie, peuvent se présenter sous des formes différentes, par exemple, AB, A ou encore EX, E, ils sont représentés par autant d'unités lexicales qu'il y a de graphies différentes :

ƀAB	AB	70600
ƀAB	A	70600

2. *Les mots à flexion irrégulière*

Ce groupe comprend les verbes anomaux (SVM, EO, FERO), quelques numéraux (DVO, TRES) et la plupart des adjectifs-pronoms. Ces mots sont le plus souvent les seuls représentants d'un type de flexion. En outre, leur radical (au sens où je l'ai défini à la page 749) est difficile, voire même impossible, à isoler (SVM, ES, EGO, MIHI).

Toutes les formes existantes de mots irréguliers sont enregistrées dans le lexique avec leur analyse complète. Par exemple :

ƀSVM	1EST	56C11
ƀEGO	MIHI	41E00

3. *Les formes susceptibles de produire de fausses analyses*

Il s'agit de formes pour lesquelles la décomposition en radical et désinence conduit à la production, à côté de l'analyse correcte, d'analyses fausses.

Ces erreurs sont dues au fait qu'il y a dans le lexique plusieurs radicaux auxquels de telles formes peuvent logiquement se rattacher.

Prenons comme exemple la forme VIRVM, accusatif singulier de VIR. Le lexique contient deux radicaux VIR :

9	VIR	VIR	123
9	VIRVS	VIR	126

Les codes 123 et 126 conduisent respectivement au choix des désinences des substantifs masculins et des substantifs neutres de la deuxième déclinaison. Par ailleurs, la désinence -VM se trouve à l'accusatif singulier des substantifs masculins et au nominatif, au vocatif et à l'accusatif des substantifs neutres de la deuxième déclinaison.

Dès lors, lorsque l'ordinateur a décomposé VIRVM en un radical VIR- et en une désinence -VM, il fournit les quatre analyses qu'il trouve dans la table des désinences, c'est-à-dire accusatif singulier de VIR, nominatif, vocatif et accusatif singuliers de VIRVS.

Pour écarter de telles erreurs, il suffit d'enregistrer VIRVM dans le lexique comme une unité-forme et d'indiquer par un code, dont je parlerai bientôt, qu'il est inutile de procéder à la décomposition de cette forme. Aucune règle ne permet de déterminer préalablement quelles sont les formes susceptibles de produire de fausses analyses. C'est l'expérience seule qui conduit à les enregistrer.

Ce troisième type d'unités lexicales peut s'appliquer à n'importe quel mot et à n'importe quelle catégorie grammaticale. Cette unité se suffit à elle-même pour la lemmatisation et l'analyse de la forme qu'elle porte puisqu'elle n'exige aucun recours à la table des désinences.

Ordonnance du lexique

Le lexique est enregistré sur disques magnétiques sous la forme d'un fichier indexé-séquentiel. Le mode de consultation du lexique, dont dépendent la souplesse et la rapidité du programme d'analyse, détermine dans une large mesure le classement des unités lexicales.

Lors de l'enregistrement, ce classement s'est fait d'abord sur l'ordre alphabétique des radicaux, puis lorsqu'il a plusieurs radicaux identiques, dans l'ordre numérique croissant des codes d'analyse. Ainsi, les substantifs se trouvent avant les adjectifs, les adjectifs avant les verbes etc. Dans l'exemple ci-dessous, nous voyons comment l'ordinateur a classé les deux radicaux AGR- :

9	AGER	AGR	123
9	AGRO	AGR	51b

Le classement par catégories grammaticales a pour seul but de faire produire les diverses analyses d'une même forme dans un ordre grammatical strict, ce qui facilite la tâche du philologue au moment du choix de l'analyse correcte. Quant à l'ordre alphabétique des radicaux, il joue un rôle capital dans la constitution et dans la consultation du lexique. On sait, en effet, que les éléments qui servent de clé dans un fichier indexé-séquentiel doivent être rangés en ordre croissant avant

même la création du fichier. Or, dans la consultation du lexique, le radical est la partie la plus importante de la clé.

Nous avons vu que, dans un fichier indexé-séquentiel, chaque clé doit être unique, c'est-à-dire qu'elle ne peut se trouver qu'une seule fois dans le fichier; la présence de radicaux identiques dans la langue interdit de faire du radical le seul élément entrant dans la constitution de la clé. Par ailleurs, au moment où l'ordinateur consulte le lexique, le radical est la seule information qui lui soit connue.

Pour distinguer les radicaux semblables, les unités lexicales classées en ordre alphabétique ont été numérotées de la façon suivante : l'ordinateur lit un radical et lui attribue l'indice 01, puis il procède à la lecture du radical suivant. Si ce dernier est identique au précédent, il reçoit l'indice 02. Cette opération se poursuit de la même façon jusqu'à ce qu'un radical différent apparaisse, de sorte que chaque radical repris dans le lexique a au moins l'indice 01. Voici un exemple de la numérotation des radicaux :

9	ALGEO	ALS	01 524
9	ALTVM	ALT	01 126
9	ALTOR	ALT	02 131
9	ALTVS	1ALT	03 143
9	ALTVS	2ALT	04 21b
9	ALØ	ALT	05 538
9	ALTAR	ALTAR	01 137

Le radical ALS est unique, il reçoit l'indice 01; les différents radicaux ALT sont numérotés de 01 à 05. ALTAR est un cas identique à ALS.

Les notes additionnelles

. Le code de PLURALIA TANTUM (- en colonne 44).

Ce code concerne les substantifs et les adjectifs qui ne sont usités qu'au pluriel. Son rôle est

d'éliminer les analyses du singulier. Prenons, par exemple, le substantif DIVITIAE, dont l'unité lexicale est :

9 DIVITIAE DIVITI 11b-

La forme DIVITIAE peut être formellement soit un génitif ou un datif singuliers, soit un nominatif ou un vocatif pluriels. Les analyses de singulier qui sont inusitées, sont rejetées par l'ordinateur grâce au tiret qui suit le sélecteur d'analyse. Ce procédé permet d'éliminer un grand nombre d'interprétations fantaisistes.

B. Les arrêts d'analyse.

I) Le code 1 en colonne 46 a pour but d'arrêter le processus d'analyse pour une forme déterminée et de provoquer le passage au traitement de la forme suivante. Ce code offre un double avantage : d'une part, il accélère le travail de l'ordinateur et, d'autre part, il permet d'éliminer certaines analyses erronées.

a.- Pour un grand nombre de formes, qu'il s'agisse de mots invariables ou de formes fléchies, on peut affirmer, sans risque d'erreurs, que la décomposition en radical et désinence n'apporte aucune analyse nouvelle. Soit, par exemple, la forme AIO que l'ordinateur analyse en utilisant dans le lexique l'unité-forme suivante :

bAIO AIO 56A11 1

Les codes 56A11 signifient que la forme AIO est un verbe anomal actif à la première personne de l'indicatif présent.

Cette analyse est la seule possible pour la forme AIO. Dès lors, cette unité lexicale porte le code 1 en colonne 46, afin d'indiquer à l'ordinateur qu'il n'y a pas lieu de procéder à la décomposition de la forme.

Cette façon d'accélérer l'analyse en arrêtant le déroulement normal du programme est également applicable lorsqu'il y a plusieurs analyses pour une même forme. Dans ce cas, il suffit que la dernière unité lexicale, c'est-à-dire celle qui a l'indice de radical le plus élevé, porte le code 1. Soit, par exemple, la forme QUOS pour laquelle on trouve trois

unités lexicales :

ƀQVI	1QVOS	01 46L	4
ƀQVIS	1QVOS	02 47L	4
ƀQVIS	2QVOS	03 48L	14

L'ordinateur découvre dans l'ordre les trois analyses suivantes pour la forme QUOS :

1) accusatif masculin pluriel de l'adjectif-pronom relatif QVI;
2) accusatif masculin pluriel de l'adjectif-pronom interrogatif QVIS;
3) accusatif masculin pluriel de l'adjectif-pronom indéfini QVIS.

Lorsqu'il propose la dernière analyse, l'ordinateur rencontre le code 1 dans l'unité lexicale. La consultation du lexique est aussitôt interrompue et le traitement passe au mot suivant.

b.- La seconde application du code 1 est d'éliminer des analyses fausses. En voici deux exemples :

1) j'ai parlé précédemment du cas de VIRVM. Cette forme est enregistrée dans le lexique avec son analyse complète et est accompagnée du code 1. De cette manière, le processus d'analyse est arrêté dès que la seule analyse correcte est produite;
2) pour la première classe d'adjectifs, il existe une série de désinences communes aux adjectifs en -VS et aux adjectifs en -ER; les seules différences entre ces deux groupes se situent au niveau du nominatif et du vocatif masculins singuliers (-VS -ER, -E -ER).

   Lorsque l'ordinateur doit analyser des adverbes de manière dérivés des adjectifs en -ER, il produit une analyse fausse. En voici la raison.

   On se souvient que les adverbes de manière sont analysés par le processus de décomposition en radical et désinence en utilisant l'unité lexicale de l'adjectif. Dans le cas des adjectifs en -ER, la désinence de l'adverbe (-E) est identique à celle du vocatif

singulier des adjectifs en -VS. Par conséquent, lorsque l'ordinateur traite la forme PIGRE, par exemple, il l'analyse non seulement comme adverbe, mais également comme vocatif singulier de PIGER. Le code 1 de la colonne 46 permet précisément d'éviter cette erreur : il suffit de faire de PIGRE une unité lexicale-forme et de l'affecter du code 1 :

bPIGRE	PIGRE	60000	1

II) Le principe général du programme d'analyse automatique est de séparer dans la forme la désinence et le radical. L'ordinateur opère cette distinction en commençant par la fin du mot et il isole successivement chaque lettre en remontant de proche en proche jusqu'à la première.

Dans la plupart des cas, ce procédé fait accomplir à l'ordinateur plusieurs opérations inutiles qui ralentissent l'analyse. Ainsi, dans le traitement de la forme CONSUETUDINEM, par exemple, la poursuite de la décomposition au-delà de CONSVETVDIN- ne conduit à aucun résultat et oblige l'ordinateur à accomplir dix lectures inutiles du lexique (CONSVETVDI-, CONSVETVD-, CONSVETV-, CONSVET-, CONSVE-, CONSV-, CONS-, CON-, CO-, C-).

Pour remédier à cet inconvénient, j'ai introduit, en colonne 46, un code 2 qui signifie pour l'ordinateur : ce radical est indécomposable. Soit la forme MULTITUDINES et l'unité lexicale :

9	MVLTITVDO	MVLTITVDIN	01133	2

Dans la forme MVLTITVDINES, l'ordinateur isole à un certain moment le radical MVLTITVDIN et la désinence -ES. Il exploite toutes les possibilités de la désinence -ES (nominatif, vocatif et accusatif pluriels), puis est arrêté dans son travail de décomposition de la forme MVLTITVDINES par la présence du code 2 dans l'unité lexicale.

Par ailleurs, ce code peut servir à l'élimination d'analyses erronées. Je prendrai comme exemple le mot DEVS.

Pour les formes DI et DIS de ce substantif, on peut considérer que le radical est D et que les désinences sont celles des substantifs pluriels de la deuxième déclinaison. L'unité lexicale est :

9 DEVS D 123-

Avant d'aller plus loin, notons que le tiret (PLURALIA TANTUM) empêche de considérer DO comme le datif-ablatif de DEVS.

La combinaison de ce radical D- avec les désinences plurielles -I, -I, -OS, ORUM, -IS, -IS des substantifs masculins de la deuxième déclinaison, entraîne une erreur pour la forme DOS, la dot, qui existe réellement dans la langue à l'inverse de la combinaison DORUM que l'ordinateur ne peut rencontrer. Pour conserver le radical D, nous avons affecté l'unité lexicale de DOS d'un code 2 qui arrête l'analyse et empêche la décomposition. Cette unité se présente de la façon suivante :

9 DOS DOS 13 22

. Notes relatives aux verbes composés.

Pour des raisons d'économie de place, les verbes composés ne font pas l'objet d'unités lexicales. En effet, leur analyse et leur lemmatisation peuvent être réalisées à partir d'un verbe simple et d'un préverbe. Ce procédé est décrit dans le chapitre consacré au programme d'analyse.

Cependant, cette façon de traiter les verbes composés influence le contenu du lexique. Les pages qui suivent ont pour but de décrire les trois caractéristiques lexicales dues au traitement des verbes composés.

1) *Les verbes simples fictifs*

Cette méthode est particulièrement économique; mais elle peut entraîner des erreurs d'analyse puisque l'ordinateur traitera comme des verbes réels les radicaux fictifs que l'on a introduit dans le lexique. Nous y avons remédié en ajoutant à ces radicaux des codes spéciaux.

Prenons le cas du verbe DO, DARE dont les composés sont formés en -DERE. Le lexique contient à côté du verbe DARE, le verbe fictif DERE, de sorte que l'on trouve les deux unités suivantes :

9	DO	D	01 51b	(DARE)
9	DO	D	02 53b	(DERE)

Lorsque l'ordinateur rencontre la forme DET, il propose deux analyses :

a) verbe DO, DARE, première conjugaison actif, troisième personne du singulier du subjonctif présent;

b) verbe DO, DERE, troisième conjugaison actif, troisième personne du singulier de l'indicatif futur.

De telles erreurs sont évitées par un code spécial (& en position 48) qui signifie que le mot n'existe qu'en composition.

2) *Les formes indécomposables*

Lorsque certains mots indécomposables commencent par un groupe de lettres identique à un préverbe, il faut empêcher l'ordinateur de décomposer ces formes. Prenons, par exemple, la forme AUDET. Le groupe AV étant un préverbe possible en latin (dans AVFERO), l'ordinateur considère AUDET comme un verbe composé du préverbe AU et de la forme DET qui peut être rattachée à DARE et à DERE. Ainsi, la forme AUDET reçoit trois analyses :

1.- AVDEO verbe semi-déponent de la deuxième conjugaison, troisième personne du singulier de l'indicatif présent;

2.- verbe AVDO première conjugaison actif, troisième personne du singulier du subjonctif présent (AV-DARE);

3.- verbe AVDO troisième conjugaison actif, troisième personne du singulier de l'indicatif futur (AV-DERE).

Ces deux dernières interprétations sont évitées par la présence d'une note additionnelle (tiret en position 48) qui signifie : ce radical ne peut donner naissance à aucune forme de verbe composé. Lorsque l'ordinateur rencontre un radical qui porte cette note, il saute la partie du programme destinée au traitement des verbes composés.

3) *Les erreurs de décomposition dues au radical*

Certaines formes de verbes composés peuvent être mal analysées parce que le radical isolé conduit à de fausses interprétations. C'est le cas, par exemple, pour la forme CONDET. L'ordinateur distingue dans CONDET, le préverbe CON et l'élément verbal DET. Dès lors, CONDET est rattaché à CONDERE et à CONDARE. Cette dernière analyse est parfaitement logique puisque DARE est un verbe simple susceptible d'entrer en composition (dans CIRCVMDARE par exemple).

Dans ce cas encore, il a fallu introduire une note additionnelle (0 en position 48) qui signifie : ce radical ne peut jamais entrer en composition. Les trois unités lexicales de DARE (thèmes du présent, du parfait et du supin) ont reçu ce code de sorte qu'elles ne peuvent jamais intervenir dans le traitement d'une forme composée. Cette solution, qui permet d'éviter le rattachement de CONDET à CONDARE qui n'existe pas, a pour conséquence indirecte d'imposer l'introduction dans le lexique de verbes composés tels que CIRCVMDARE.

D. Il reste à examiner une note dont le rôle est davantage technique : le tiret en dernière position de l'enregistrement.

Lorsque l'ordinateur a isolé un radical, il consulte le lexique et trouve normalement l'unité lexicale dont il a besoin. Il va ensuite chercher, dans la table des désinences, les éléments qui lui permettent de fournir la ou les analyses du mot; le cycle doit se répéter aussi longtemps que toutes les unités lexicales identiques au radical n'ont pas été exploitées.

Après examen des diverses possibilités, l'ordinateur doit en principe continuer la décomposition en remontant chaque fois d'une lettre vers la première lettre du mot. Le tiret que contient la colonne 49 indique à l'ordinateur le moment où il doit précisément procéder à une nouvelle décomposition;

il a l'avantage d'épargner pour chaque radical une lecture inutile du fichier et contribue dès lors à accélérer le traitement.

Soit par exemple, à analyser la forme DEMENTIA.

Nous trouvons, dans le lexique, les trois radicaux suivants :

9	DEMENS	DEMENT	0125	-
9	DEMENTIA	DEMENTI	0111	
9	DEMENTIO	DEMENTI	0254	-

La première décomposition de la forme DEMENTIA donnera un radical DEMENTI et une désinence A.

La première unité DEMENTI est susceptible de fournir trois analyses (nominatif, vocatif et ablatif singuliers de DEMENTIA).

L'ordinateur passe ensuite à l'exploitation de la deuxième unité DEMENTI pour laquelle n'existe aucune désinence A. Cette seconde unité porte une information qui va diriger la suite du travail : il s'agit du tiret dans la dernière position de l'enregistrement. Ce dernier signifie qu'il n'existe pas d'autre radical DEMENTI et qu'une nouvelle décomposition doit être effectuée.

On obtient ensuite le radical DEMENT- et la désinence -IA pour lesquels d'autres analyses sont découvertes.

Au terme de cette analyse du lexique, je rappellerai en quelques mots les caractéristiques qui viennent d'être énoncées.

1) Chaque unité lexicale appartient à un type déterminé par son mode d'utilisation, unité-forme pour les mots invariables ou irréguliers, unité-radical pour les mots à flexion régulière et unité-générateur pour les mots dont on ne peut former le nominatif à partir du radical.

2) Les adverbes de manière ne figurent pas dans le lexique; c'est l'unité lexicale de l'adjectif à partir duquel ils sont régulièrement formés qui sert à leur analyse. De la même manière, les verbes composés ont été omis lorsqu'ils peuvent s'analyser sans risque d'erreur à partir d'un préverbe et d'un verbe simple.

3) Un jeu de notes additionnelles guide le processus d'analyse et de lemmatisation. Elles sont destinées à éviter des erreurs et à accélérer le cycle des opérations.

Contenu du lexique

Dès l'origine du traitement automatique du latin, un double principe a guidé la constitution du lexique. Seuls les mots qui se trouvaient dans des textes déjà lemmatisés et analysés ont été repris dans le lexique à condition qu'ils soient employés un certain nombre de fois dans l'ensemble de la littérature latine. Cette façon purement empirique de déterminer le contenu lexical exige une mise à jour continuelle du dictionnaire de l'ordinateur, sous forme d'un enrichissement progressif au fur et à mesure que des textes sont traités. En outre, lorsque l'ordinateur doit analyser des oeuvres qui contiennent des mots particulièrement liés au sujet, il est toujours possible d'introduire provisoirement ces mots dans le lexique afin d'en augmenter la rentabilité.

Ce mode de constitution du lexique présente un inconvénient : l'ordinateur ne peut analyser toutes les formes d'un texte et, dans quelques cas, il ne propose que des solutions fausses. C'est là un inconvénient mineur puisque, de toute manière, le philologue vérifie et corrige le travail de l'ordinateur. Au surplus, l'enrichissement continuel du lexique élimine de plus en plus les risques d'erreur.

Actuellement, le lexique de l'ordinateur se compose de 14.200 unités qui représentent 7.215 mots différents. Au point de vue des catégories grammaticales, ces mots se répartissent comme suit :

Catégorie	Effectifs	Pourcentage
Substantifs	3.408	47,24
Adjectifs	1.633	22,63
Numéraux	74	1,03
Adjectifs-pronoms	98	1,36
Verbes	1.343	18,61
Adverbes	521	7,22
Prépositions	44	0,61
Conjonctions	76	1,05
Interjections	18	0,25
TOTAL	7.215	100, --

Ce tableau appelle quelques remarques. Tout d'abord, le nombre de verbes est relativement faible par rapport au nombre des substantifs et des adjectifs. Ce phénomène résulte du fait que peu de verbes composés sont repris dans le lexique. Si l'on veut avoir une idée plus précise du pourcentage de lemmes de verbes dans la langue, il faut procéder à un dénombrement des verbes composés. La même remarque peut être faite à propos des adverbes puisque les adverbes de manière ne font pas l'objet d'unités lexicales.

Enfin, on peut considérer que pour les catégories de mots invariables et pour les adjectifs-pronoms, le lexique ne subira plus guère de modifications; en effet, la quantité de textes sur laquelle se fonde notre lexique permet de considérer que presque tous les mots invariables qui existent en latin ont été rencontrés dans un texte et introduits dans le lexique.

CHAPITRE IV

La table des désinences

Après avoir décomposé un mot en radical et terminaison et après avoir découvert une unité lexicale compatible avec le radical isolé, l'ordinateur consulte la table des désinences pour y découvrir l'analyse du mot. Cela signifie que ce n'est pas la forme qui est analysée, mais plutôt la désinence.

Dès lors, la table des désinences contient obligatoirement toutes les analyses morphologiques possibles et, parallèlement, toutes les terminaisons des mots à flexion; c'est pour cette raison que les désinences formellement identiques figurent dans la table autant de fois qu'elles sont susceptibles de recevoir d'analyses différentes.

En procédant de la sorte, nous avons dénombré 3.068 désinences réparties en 62 types différents. Chaque catégorie correspond à un des 62 sélecteurs d'analyse dont la liste figure dans le chapitre consacré au lexique.

La longueur d'une unité de désinence est de vingt-deux positions et son schéma est le suivant :

Col. 1 - 3	sélecteur d'analyse
Col. 4 - 5	indice de désinence
Col. 6 -14	désinence
Col. 15	-
Col. 16	code de fin de groupe
Col. 17-22	analyse

Voici un exemple :

11b01	A - 11A00

Les colonnes 1 à 3 sont rigoureusement identiques aux colonnes 41 à 43 des unités lexicales avec radical. Cette identité est fondamentale dans le processus d'analyse puisque le rôle du sélecteur est

de préciser la série de désinences qui est compatible avec chaque radical.

Les deux chiffres en position 4 et 5 - 01 - jouent le même rôle que l'indice du radical dont j'ai parlé précédemment : ils différencient les désinences identiques.

La désinence occupe les colonnes 6 à 14 et elle se trouve toujours alignée sur la droite; ainsi, dans l'exemple de la page précédente, A se trouve en colonne 14.

Le tiret est la seule information que l'on puisse trouver en colonne 16; il sert à arrêter la consultation de la table pour une désinence déterminée. Il agit de la même façon que le tiret qui se trouve à la position 49 des unités lexicales. Enfin, les colonnes 17 à 22 portent l'analyse morphologique complète de la désinence. La colonne 17 correspond à la colonne 54 du code du Laboratoire d'Analyse statistique des Langues anciennes et la colonne 21 à la colonne 58. La colonne 22 est réservée au genre des mots autres que les substantifs.

Ordonnance de la table et mode de consultation

Contrairement aux unités lexicales, les désinences sont classées selon un ordre extrêmement complexe.

Trois critères interviennent dans le classement.

1.- Le sélecteur d'analyse

Les désinences sont rangées en ordre croissant des sélecteurs d'analyse par catégories et sous-catégories grammaticales. Ainsi, les désinences des substantifs de la première déclinaison sont rangées en tête de la table; elles sont suivies des désinences des substantifs de la deuxième déclinaison.

2.- L'indice de désinence

Pour un même sélecteur, il peut exister plusieurs désinences identiques; celles-ci sont numérotées (01, 02 ..) selon le même procédé que les radicaux dans le lexique : chaque désinence reprise dans la table a au moins l'indice 01. Ainsi, pour la première déclinaison, on trouve au début de

la série de désinences tous les sélecteurs numérotés 01; on trouve également une désinence -IS marquée 02 puisqu'il y a deux analyses possibles pour cette désinence (datif et ablatif pluriels).

3.- Ordre alphabétique des désinences

A l'intérieur d'un groupe de désinences portant le même indice, les unités sont classées en ordre alphabétique. Ainsi, pour la première déclinaison, la désinence -A est suivie de -AE, -AM, -AS, -IS, -ARVM.

En ce qui concerne la consultation de la table des désinences, plusieurs méthodes étaient possibles. J'en citerai trois.

Pour illustrer ce qui suit, je prendrai comme exemple la forme AMICITIAE que l'ordinateur a décomposée en un radical AMICITI- et en une désinence -AE.

1.- Partant du groupe AE, l'ordinateur peut lire les 3.068 unités de la table pour y chercher les quatre désinences AE compatibles avec un substantif de la première déclinaison. Cette solution est évidemment trop lente puisqu'elle demande une lecture intégrale de la table pour chaque forme à analyser.

2.- La deuxième solution à laquelle on aurait pu recourir était de consulter, en partant du sélecteur d'analyse fourni par le lexique, toutes les désinences possibles pour ce sélecteur. Ce système est relativement économique pour les catégories qui n'ont que quelques terminaisons, mais pour les verbes, par exemple, il impose de lire chaque fois 210 unités-désinences.

Au surplus, pour utiliser ce procédé, il est nécessaire de trouver un moyen rapide d'accéder directement à chaque type de sélecteur, ce qui techniquement est difficile à réaliser.

3.- La méthode utilisée pour la consultation de la table des désinences fait appel au mode indexé-séquentiel de lecture des fichiers. Ce procédé, on le sait, repose sur la constitution d'une clé de lecture qui permet d'identifier à coup sûr ce que l'on cherche.

Les éléments qui interviennent dans la constitution de la clé sont :

a) le sélecteur d'analyse fourni par l'unité lexicale. Ainsi, AMICITI- est accompagné du sélecteur 11b;

b) un indice auquel l'ordinateur attribue toujours la valeur 01 au départ d'une consultation de la table. Cet indice augmente de 1 jusqu'à ce qu'une unité-désinence portant un tiret en colonne 16 ait été lue;

c) la lettre ou le groupe de lettres que l'ordinateur a détaché de la forme pour isoler le radical. Dans notre exemple, on aura -AE.

Pour AMICITIAE, l'ordinateur va construire quatre clés qui fourniront chacune une analyse.

clé		analyse	
11b01	AE	11D00	(génitif singulier)
11b02	AE	11E00	(datif singulier)
11b03	AE	11J00	(nominatif pluriel)
11b04	AE	11K00	(vocatif pluriel)

La dernière unité porte un tiret en position 16, ce qui signifie : ceci est la dernière désinence AE possible pour un substantif de la première déclinaison. Le processus de consultation de la table est aussitôt arrêté. Ce procédé présente, par rapport aux précédents, des avantages certains : il est plus rapide puisqu'il se limite à la lecture des unités qui sont compatibles avec le radical et avec la désinence isolée. Un second avantage réside dans le fait que chaque lecture dans la table des désinences fournit une analyse correcte sans qu'il soit besoin de vérifier si l'analyse trouvée est compatible avec le radical. En effet, s'il n'y a pas d'analyse pour une terminaison, la lecture du fichier indexé-séquentiel est impossible.

Cas particuliers

1) Parmi les nombreuses distinctions que nous avons faites pour la troisième déclinaison, il faut signaler le cas des substantifs en -O, -ONIS, AS, ATIS et OR, ORIS. Trois séries spéciales de désinences ont été créées pour ces substantifs. Ainsi, on trouve les radicaux MONITI (MONITIO), AM (AMOR) et AVCTORIT (AUCTORITAS) auxquels peuvent s'adjoindre respectivement les désinences -O, -ONIS,

ONEM ..., -OR, ORIS, OREM ..., -AS, ATIS, ATEM ... En outre, ces radicaux ont tous reçu le même sélecteur, à savoir 13b. Cependant, il est très vite apparu que cette classification donnait lieu à des analyses fausses; en voici un exemple :

pour la forme GENITAS, l'ordinateur isole un radical GENIT- et une désinence -AS.
Il y a dans le lexique, deux radicaux GENIT- :

9	GENITOR	GENIT	0113
9	GIGNO	GENIT	02538

La forme GENITAS a comme seule interprétation possible : participe parfait passif accusatif féminin pluriel de GIGNO et cette interprétation doit être produite à partir de la seconde unité GENIT. Mais la solution qui consiste à mélanger les désinences des substantifs en -OR et en -AS de la troisième déclinaison et à leur donner le même sélecteur conduit à deux analyses erronées qui sont : nominatif et vocatif singuliers de GENITOR.

En effet, l'ordinateur découvre dans le lexique un radical GENIT qui est compatible avec toutes les désinences caractérisées par le code 13. D'autre part, il a isolé une terminaison -AS qui lui permet de créer la clé de lecture : 13b01 AS pour laquelle il trouve effectivement les unités-désinences 13b01 AS 13A00 (nominatif)
13b02 AS 13B00 (vocatif)

Il va de soi que les sélecteurs de désinences ont dû être modifiés tant dans le lexique que dans la table des désinences, de façon à distinguer les substantifs en -O, en -OR et en -AS.

Les substantifs en -O, ONIS sont caractérisés par le sélecteur 13b; ceux en -OR, ORIS par le sélecteur 131 et ceux en -AS, ATIS par le sélecteur 132.

Exemple :	9	MONITIO	MONITI	01 13b
	9	GENITOR	GENIT	01 131
	9	AUCTORITAS	AUCTORIT	01 132

Les désinences également modifiées sont rangées dans des catégories distinctes, de sorte que -AS figure dans les désinences 132 et non plus dans les désinences 13b.

Ainsi, pour GENIT -AS, l'ordinateur continue de découvrir le radical GENIT- de GENITOR, mais la clé de désinences qu'il crée n'est plus 136 AS mais bien 131 AS et comme on ne peut trouver les désinences -AS qu'avec la clé 132 AS, la clé 131 AS ne conduit à aucune analyse.

2) Lorsque la forme est identique au radical, on a affaire à une désinence zéro. Ainsi, TIME est à la fois le radical et la deuxième personne du singulier de l'impératif présent de TIMEO. La désinence zéro ne présente formellement aucune particularité et il est possible de la traiter selon le procédé normal; c'est pourquoi on trouve quelques exemples de la désinence zéro : DIE, AUDI, TIME.
Il n'est pas impossible d'étendre l'utilisation de la désinence zéro à certains substantifs de la troisième déclinaison (CONSUL), mais jusqu'à présent nous avons considéré qu'il s'agissait d'une amélioration de détail qui ne modifie nullement les performances du programme d'analyse.

3) L'existence de désinences ayant la même valeur mais des formes différentes (-ERE, -ERUNT pour le parfait) ne pose guère de problèmes. Il suffit que ces deux terminaisons figurent dans la table avec la même analyse.

Contenu de la table

A l'opposé du lexique dont le contenu a été fixé empiriquement, la table des désinences doit contenir toutes les désinences de la langue. Toutefois, nous avons exclu quelques désinences rares propres à multiplier inutilement le nombre d'analyses par forme. C'est le cas notamment pour la finale -IS, qui peut se trouver à l'accusatif pluriel de la troisième déclinaison, ainsi que pour -OS nominatif singulier de la deuxième déclinaison. En effet, à supposer que -OS nominatif se trouve dans le lexique, on aurait pour chaque forme d'accusatif pluriel, deux analyses dont l'une est exceptionnelle.

Je terminerai ce chapitre consacré aux désinences par un résumé en deux points :

1.- la table des désinences contient toutes les terminaisons des catégories grammaticales à flexion régulière à l'exception de quelques désinences rares.

L'existence d'une désinence zéro n'entrave nullement le travail de l'ordinateur;

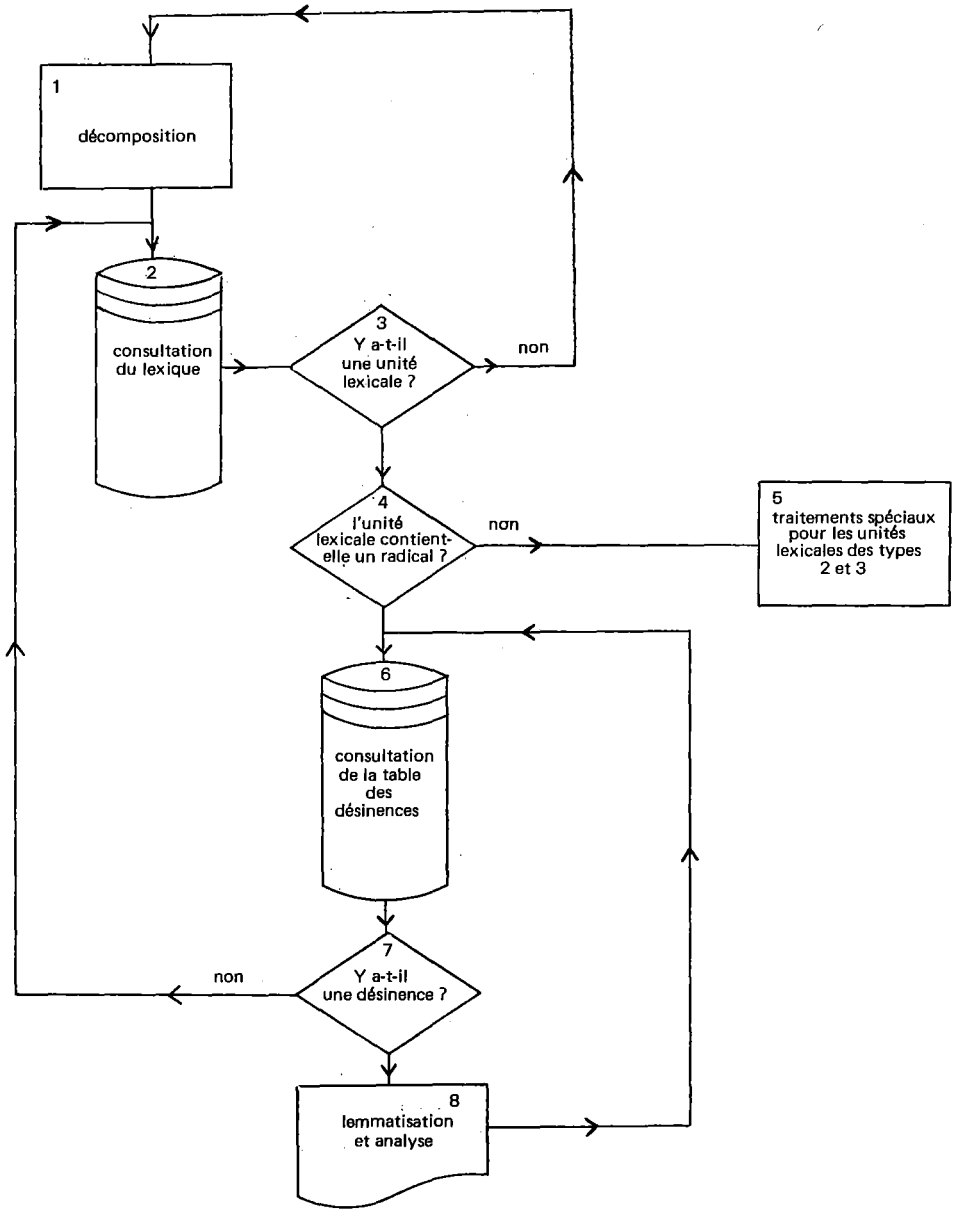
2.- le classement des désinences dans la table et le mode de consultation permettent de donner au programme un maximum d'efficacité et de sécurité puisque chaque lecture de la table fournit nécessairement une analyse formellement possible sans que l'ordinateur procède à une vérification.

CHAPITRE V

Le programme d'analyse automatique

Pour produire le lemme et l'analyse d'un mot latin, le programme exécute essentiellement deux opérations : la décomposition de la forme et la consultation du lexique et de la table des désinences.

La décomposition de la forme s'effectue à partir de la fin du mot en remontant progressivement lettre par lettre vers le début. Chaque nouvelle décomposition est suivie d'une consultation du lexique et éventuellement d'une consultation de la table des désinences. Si l'on tient compte uniquement des unités lexicales des mots à flexion (unités-radical), le schéma simplifié des opérations est le suivant :



1) Décomposition de la forme à partir de la droite en radical et désinence possibles.

2) Consultation du lexique.

3) S'il n'y a pas ou s'il n'y a plus d'unité lexicale correspondant au groupe de lettres isolées, l'ordinateur répète le processus de décomposition (opération 1).

4) Si l'unité lexicale contient un radical, l'ordinateur exécute l'opération 6.

5) Un traitement simplifié est prévu pour les mots irréguliers et invariables et pour les formes à analyser par une unité lexicale du deuxième type.

6) Consultation de la table des désinences.

7) S'il n'y a pas ou s'il n'y a plus de désinence compatible avec le radical, l'ordinateur consulte à nouveau le lexique.

8) Enregistrement d'un lemme et d'une analyse et retour à la consultation de la table des désinences.

Cette méthode de décomposition des formes et de consultation du lexique et des désinences diffère des deux procédés employés auparavant.

Le premier procédé a été utilisé par le Laboratoire d'Analyse statistique des Langues anciennes pour le premier programme d'analyse du latin. Dans ce système, le traitement d'une forme se fait de la façon suivante :

1.- décomposition de la forme à partir de la droite;

2.- recherche dans la table des désinences d'une unité correspondant aux éléments isolés;

3.- consultation du lexique des radicaux.

Le second système qui a été proposé préconise la décomposition de la forme à partir de la gauche, c'est-à-dire à partir du début du mot. La consultation des lexiques se fait également à partir de l'élément de gauche, c'est-à-dire à partir du radical supposé. Le principe qui sert de base théorique à cette méthode est le suivant :

"les avantages de la reconnaissance des formes à partir de la gauche sont liés au fait que la machine consulte d'abord le lexique des formes, en commençant par la gauche, et non pas des listes fermées

de désinences, en commençant par la droite" (11).

Le but de ce procédé est de "diminuer le nombre d'hypothèses aberrantes" et son "premier avantage, à condition ... de ne donner ni des unités lexicales précédent la parenthèse de moins de quatre lettres ni des préfixes, réside dans le fait que la machine ira plus vite, parce qu'elle créera beaucoup moins d'hypothèses inutiles" (12).

Cette méthode est inacceptable et relève de l'utopie : comment faut-il traiter le nombre considérable de mots qui ont un radical de moins de quatre lettres ? Il est certain, par exemple, qu'en latin ce procédé est inapplicable puisque la longueur moyenne du mot, désinence incluse, se situe aux environs de six lettres et que, si l'on retire les désinences, cette longueur est le plus souvent inférieure à quatre lettres.

Ce qui est plus grave encore dans cette méthode, c'est qu'il y a une confusion évidente entre la décomposition de la forme, d'une part, et la consultation des lexiques, d'autre part.

La différence entre le premier programme d'analyse du latin et le système que j'ai réalisé se situe dans l'ordre des opérations 2 et 3. Dans l'ancien programme, la décomposition et la consultation des lexiques se font à partir de la droite, tandis que ma méthode consiste à décomposer le mot à partir de la droite et à consulter les lexiques à partir de l'élément de gauche, c'est-à-dire à partir du radical.

Pour montrer les avantages du système de décomposition par la droite et de consultation par la gauche par rapport à l'ancien programme d'analyse du latin, j'examinerai les différentes opérations que l'ordinateur doit accomplir pour analyser la forme ERRANT.

1) Consultation par la gauche

ERRANT

Radical	Désinence	Lex. Dés.	Lex. Rad.
ERRAN	-T	1) T existe	2) ERRAN n'existe pas
ERRA	-NT	3) NT existe	4) ERRA n'existe pas
ERR	-ANT	5) ANT existe	6) ERR existe

2) Consultation par la droite

ERRANT

Radical	Désinence	Lex. Rad.		Lex. Dés.	
ERRAN	-T	1) ERRAN	n'existe pas		-
ERRA	-NT	2) ERRA	n'existe pas		-
ERR	-ANT	3) ERR	existe	4) ANT	existe

Dans la première version du programme d'analyse du latin, l'ordinateur effectuait six consultations des lexiques pour l'analyse de la forme ERRANT; actuellement, quatre consultations suffisent. Cette différence s'accroît en proportion de la longueur de la désinence. Ainsi, pour la forme AMABAT, le nombre de consultation est huit pour l'ancien programme, contre cinq actuellement. En effet, les désinences T, AT, BAT existant dans la langue et ne pouvant convenir à la forme AMABAT, conduisent l'ordinateur à trois consultations inutiles du lexique des radicaux. Par contre, en examinant d'abord le lexique des radicaux, on n'aboutit qu'à trois consultations improductives.

Après avoir exposé le processus général de décomposition, j'examinerai à présent les grandes lignes du déroulement des opérations.

A. *Lecture des formes*

Les données d'entrée qu'utilise l'ordinateur sont les formes d'un texte latin quelconque classées en ordre alphabétique et stockées sur disques magnétiques. La mise en ordre alphabétique n'est pas obligatoire; il est possible d'analyser les formes dans l'ordre naturel du texte.

C'est ainsi que procédait le premier programme, mais ce mode de travail, plus lent, conduit à d'importantes pertes de temps.

En effet, partant de l'ordre alphabétique intégral, l'ordinateur analyse la première forme qu'il

rencontre et reproduit ensuite la ou les analyses autant de fois qu'il y a de formes identiques dans le texte. Ainsi, si la forme EST se trouve cent fois dans un texte, l'ordinateur procède une seule fois à la recherche de son analyse puis il la recopie cent fois à très grande vitesse sans consulter le lexique. L'économie de temps que l'on réalise ainsi est considérable puisque, pour un texte de 10.000 mots, ce procédé réduit presque de moitié le nombre de formes à analyser. Il est clair que ce mode de traitement est particulièrement intéressant lorsqu'on analyse en une seule fois une quantité suffisante de texte. L'expérience montre que la durée des opérations reste à peu près constante pour un texte de 9.000 mots ou pour un texte de 12.000. La raison en est que le nombre de formes nouvelles dans un texte tend à diminuer dès que l'on a parcouru une quantité suffisante de ce texte.

B. *Préparation des données*

Après la phase de lecture, les données subissent éventuellement une légère transformation. La forme enregistrée respecte rigoureusement l'orthographe de l'édition, mais il a semblé préférable d'adopter l'orthographe ancienne dans le lexique et dans la table des désinences : on ne trouvera donc pas dans ceux-ci les lettres U et J qui sont remplacées respectivement par V et I. Par conséquent, en vue de la consultation de ces fichiers, la forme doit être transportée dans une zone spéciale de la mémoire et subir les transformations nécessaires.

C. *Décomposition des formes et consultation du lexique et de la table des désinences*

Le processus de décomposition des formes et la méthode de lecture du lexique et de la table des désinences ont été exposés au début de ce chapitre. Il reste à voir comment l'ordinateur traite différemment les trois types d'unités lexicales.

Deux zones-mémoire sont nécessaires pour effectuer les coupes successives que subit une forme; l'une contient le radical et l'autre la désinence. Prenons par exemple la forme AMAT et voyons ce que vont contenir les deux zones.

	zone-radical	zone-désinence
1	AMAT	ƀ
2	AMA	T
3	AM	AT

Après la première décomposition, l'ordinateur recherche dans le lexique la forme entière (AMAT). Il y a, à cela, plusieurs raisons.

D'une part, c'est à partir de la forme entière que l'on traite les mots invariables (NON) et les mots irréguliers (SUM) qui sont représentés dans le lexique par des unités-forme. C'est également à partir de la forme entière que sont analysés les mots pour lesquels existe une unité lexicale à générateur.

D'autre part, dans le cas des mots à désinence zéro pour lesquels la forme est identique au radical (MONE est à la fois l'impératif présent actif deuxième personne du singulier et le radical de MONEO), la recherche de l'analyse doit se faire à partir de la forme entière. Dans ce cas, la désinence est représentée dans la table par une position blanche identique au contenu de la zone-désinence au début du traitement.

A partir de la deuxième décomposition (AMA), le traitement par les unités-forme et par les unités-générateur devient impossible, seule l'unité-radical peut encore intervenir dans l'analyse d'une forme.

Après chaque consultation du lexique et de la table des désinences, l'ordinateur reçoit des indications qui guident la suite du traitement. Ces indications sont constituées par les notes additionnelles que contiennent le lexique et la table des désinences. Le moment où l'ordinateur détecte ces indicateurs a une grande importance puisque, selon les cas, ils peuvent signifier, par exemple, que toutes les possibilités ont été examinées pour une forme ou seulement pour un radical ou encore pour une désinence.

D. Il reste pour terminer cette description du programme à mentionner deux boucles de programme prévues pour le traitement des adverbes de manière et pour la décomposition des verbes composés.

Pour les adverbes de manière, le programme tire parti du fait qu'ils peuvent se rattacher à la flexion des adjectifs. Dès lors, si dans le processus de décomposition l'ordinateur découvre, d'une part, un

radical d'adjectif et, d'autre part, une finale identique à un suffixe d'adverbe, une boucle de programme relativement brève permet de reconstituer le lemme adverbial à partir du radical de l'adjectif. Prenons, par exemple, la forme FELICIVS. L'ordinateur isole le radical FELIC qui se rattache à l'unité lexicale de l'adjectif FELIX et identifie IVS comme désinence adverbiale. Dès lors, le suffixe ITER est accolé au radical FELIC de manière à former le lemme FELICITER.

La lemmatisation et l'analyse d'une forme verbale composée peut s'effectuer à partir de trois éléments qui sont le préverbe, le radical d'un verbe simple et une désinence. L'exploitation de cette possibilité permet l'économie d'un grand nombre d'unités lexicales.

Pour les verbes composés, le processus de décomposition est double, il se fait simultanément par la fin du mot pour isoler la désinence et par le début pour séparer le préverbe. L'ordinateur dispose à cet effet d'une table de préverbes à laquelle il compare la lettre ou les lettres qu'il a isolées au début du mot. Si l'élément isolé est identique à l'un des préverbes, le processus normal d'analyse se déroule lorsqu'on a éliminé, pour la consultation du lexique, la partie du mot considérée comme un préverbe. C'est donc le radical du verbe simple qui sert ici à l'analyse et à la lemmatisation.

En ce qui concerne l'analyse d'une forme de verbe composé, elle est la même que celle de la forme du verbe simple : ADAMAS doit être analysé de la même façon que AMAS et l'ordinateur découvrira cette analyse en décomposant ADAMAS en AD - AM - AS.

Pour former le lemme d'un verbe composé, l'ordinateur place le préverbe (AD) devant le lemme du verbe simple (AMO).

Le programme de lemmatisation et d'analyse morphologique du latin est au centre du système dont j'ai parlé précédemment, tant par sa situation que par son rôle; c'est lui qui réalise la tâche la plus complexe.

Cependant, la description qui précède fait apparaître la simplicité du programme dont le rôle consiste, en définitive, à découper un mot de manière systématique et à consulter après chaque décomposition des fichiers enregistrés sur disques magnétiques. L'ordinateur ne procède à aucun arrangement des

données et l'unique transformation qu'il opère porte sur les seuls adverbes de manière. La suppression de cette boucle enlève au programme toute particularité propre au latin et permet la création d'un programme d'analyse morphologique de n'importe quelle langue flexionnelle. Ce programme serait capable d'utiliser différents lexiques constitués sur un schéma identique, capable aussi d'analyser diverses langues, par exemple, le grec et le français. Ce résultat n'a pu être obtenu qu'en considérant les mots d'un texte comme des éléments formels totalement indépendant du contexte. Je donnerai, à titre d'exemple, dans la suite de cet exposé, une description d'un lexique du français.

Il est une autre conséquence de la souplesse et de la simplicité du programme : le rôle le plus important dans l'analyse morphologique est dévolu au lexique et à la table des désinences. Ce sont ces fichiers qui contiennent toutes les réponses aux questions que l'on pose à l'ordinateur et tous les éléments qui permettent d'organiser les réponses.

Performance du système d'analyse automatique

Pour terminer cette description du programme d'analyse du latin, je dirai quelques mots de son efficacité.

Je prendrai comme exemple pour illustrer ce point les mille premiers mots des *Géorgiques* de VIRGILE. Ce texte a été choisi parce qu'il est réputé difficile et parce qu'il contient un grand nombre de noms propres qui ne se trouvent pas dans le lexique. En outre, il faut souligner que, jusqu'à présent, notre dictionnaire a surtout été utilisé pour des oeuvres de prose où ses performances sont bien supérieures à celles que l'on observe en poésie.

Par conséquent, les chiffres que l'on trouvera ici montrent l'efficacité du lexique dans les conditions les moins favorables.

Sur 1.000 mots :	Formes non lemmatisées et non analysées	Formes non analysées	Total (sans) noms propres	Total général
Substantifs	52	8	32	60
Adjectifs	14	5	13	19
Numéraux	-	-	-	-
Adjectifs-pronoms	-	1	1	1
Verbes	4	7	11	11
Adverbes	-	-	-	-
Conjonctions	-	-	-	-
Interjections	-	-	-	-
TOTAL	70	21	57	91

Sur un total de 1.000 formes, 91 ne sont pas traitées par l'ordinateur. Cela signifie que, pour un texte difficile, l'efficacité du programme se situe à peu près à 91 %.

Il faut encore signaler que, par rapport à des textes de prose, deux particularités diminuent l'efficacité du programme. C'est tout d'abord le nombre élevé de noms propres et d'adjectifs dérivés de noms propres : il y en a 34 au total. La seconde différence s'observe dans le nombre de substantifs et d'adjectifs pour lesquels l'ordinateur propose un lemme correct sans en découvrir l'analyse. Les treize formes pour lesquelles ce phénomène se produit sont toutes des accusatifs pluriels en -IS, désinence qui n'est pas reprise dans la table.

Ainsi, sur les 91 formes que l'ordinateur n'identifie pas, 47 ont peu de chance de se trouver dans un texte en prose; les performances du programme seraient considérablement accrues si l'on faisait abstraction de ces noms propres et de la désinence -IS puisqu'on arriverait à 95 % d'analyses et de lemmes corrects.

Quoiqu'il en soit, on peut dire qu'avec un lexique de moins de 7.500 mots, l'ordinateur parvient à identifier plus de 90 % des oeuvres, qu'elles appartiennent à la prose ou à la poésie.

Il est certain qu'un fort accroissement du nombre d'articles du dictionnaire-machine augmenterait à peine les performances du programme.

CHAPITRE VI

Le problème de l'amphibologie

Le caractère amphibologique de la langue latine conduit l'ordinateur à produire pour une forme plusieurs analyses possibles.

Ainsi, l'expérience a montré que pour la plupart des textes, sinon tous, l'ordinateur fournit en moyenne un peu moins de trois analyses par forme : il est fréquent qu'une forme reçoive cinq ou six analyses différentes voire même davantage (la forme NATURA, par exemple, reçoit dix-sept analyses différentes). Cela signifie que pour une oeuvre de 10.000 mots, l'ordinateur produit environ 28.000 analyses parmi lesquelles le philologue devra choisir. Un bon exemple est donné par les 250 premiers mots des *Géorgiques* pour lesquels l'ordinateur fournit 720 analyses différentes.

Tous les essais effectués jusqu'à présent en vue de pousser plus avant l'automatisation de l'analyse n'ont abouti à aucun résultat définitif, pas plus en latin que dans les autres langues.

En vérité, il n'y a eu jusqu'à présent que des ébauches de solutions; aucune recherche systématique n'a jamais été entreprise et les rares études qui ont été faites sont restées sur un plan purement théorique sans que des essais réels en ordinateur aient été effectués (13).

On peut penser cependant que la possibilité est réelle de voir se développer progressivement des systèmes qui permettront de lever un certain nombre d'ambiguïtés. Bien sûr, dans un premier temps, l'ordinateur ne pourra tout résoudre et certains risques d'erreurs subsisteront, mais c'est précisément l'étude de ces erreurs, de ces amphibologies réelles, qui devrait conduire à la découverte de solutions nouvelles.

Dans les pages qui suivent, je voudrais tenter de montrer comment on pourrait donner à l'ordinateur quelques critères formels qui lui permettraient le choix de l'analyse correcte.

Le système qu'il faudrait réaliser comporterait au moins deux phases principales; la première phase serait constituée par le programme d'analyse morphologique tel qu'il vient d'être décrit dans les

chapitres précédents : l'ordinateur propose pour chaque mot toutes les possibilités d'analyses formelles. La seconde phase du travail consisterait à faire lever les amphibologies par l'ordinateur en lui faisant utiliser comme données d'entrée, les résultats obtenus au cours du premier stade.

Il est évident que dans un texte quelconque se trouvent des mots pour lesquels l'ordinateur ne propose qu'une seule analyse. C'est ainsi que, pour les deux cent cinquante mots analysés par l'ordinateur au début des *Géorgiques*, 95 formes n'ont qu'un seul lemme et une seule analyse, c'est-à-dire que pour trente-huit pourcents du texte, l'interprétation est claire. Ces formes à analyse unique ne pourraient-elles servir de points de repère à l'ordinateur et lui permettre de découvrir l'analyse des mots amphibologiques qui se trouvent dans le contexte immédiat ?

Un tel critère semble bien être utilisable pour certaines langues telles que le français, par exemple. C'est ainsi que les règles de levée des ambiguïtés énoncées par J. BLOIS et J. BUYDENS dans l'article auquel j'ai déjà fait allusion, se fondent sur la place des mots dans des contextes réduits ainsi que sur la présence d'un élément dont l'analyse est certaine. Ainsi, la règle nº 20 (14) dit : "Un ambigu, pouvant être adjectif qualificatif, placé entre un déterminatif non ambigu et un nom, ne peut être qu'adjectif. Exemple ... du *ferme* courage ... ".

Semblables règles sont inapplicables au latin soit parce que cette langue est dépourvue d'articles, soit encore parce que l'ordre des mots n'y est pas fixe; c'est pourquoi on peut dire : "Ainsi donc, même si l'on s'efforce de faire intervenir des règles formelles de compatibilité ou de formation de petits groupes de mots, il ne semble guère possible d'arriver à des résultats intéressants à moins que de se référer à la signification globale de la phrase et même du passage. Et à cela, je ne crois pas qu'on y arrivera de sitôt avec l'ordinateur." (15).

Pour illustrer les difficultés de la levée automatique des ambiguïtés, je prendrai comme exemple les prépositions et leur régime. Il s'agit d'un groupe qui, en apparence, est simple mais on verra que, même à ce niveau, les causes d'amphibologies sont multiples.

L'analyse d'une forme de préposition comprend deux informations : la catégorie grammaticale (code 7 en colonne 54) et le cas régi par cette préposition (colonne 56). En latin, le mot régi par une préposition peut être à l'accusatif (code 3), à l'ablatif (code 6) ou, plus rarement, au génitif (code 4). Le lexique automatique contient quarante-quatre prépositions différentes qui sont représentées par

soixante unités lexicales. L'existence de plusieurs unités pour un même lemme résulte soit de la polymorphie (AB, A; APUD, APUT), soit du fait qu'une même préposition peut régir des cas différents : ainsi, le mot régi par IN est soit à l'accusatif, soit à l'ablatif, le mot régi par TENUS est soit au génitif, soit à l'ablatif.

Avant d'examiner les cas d'ambiguïtés qui se rencontrent dans un groupe préposition-mot régi, il convient de classer les prépositions selon deux types différents.

1) Le mot que l'ordinateur traite ne peut être qu'une préposition. Mais c'est :

   - soit une préposition qui régit un cas bien déterminé, toujours le même. La forme ne prête à aucune ambiguïté possible.
     Il y a treize formes qui se comportent ainsi; ce sont : AB, AD, APUD, APUT, DE, ERGA, EX, E, INTER, OB, PENES, PER, TRANS;
   - soit une préposition qui régit l'accusatif ou l'ablatif. C'est à ce groupe qu'appartiennent IN et SUB.

2) Le mot à traiter est amphibologique dans sa forme. Il peut être une préposition ou un mot homographe d'une autre catégorie grammaticale. S'il est préposition, il peut :

   - soit régir un cas bien déterminé. Ainsi, PRO est ou une interjection ou une préposition dont le régime est à l'ablatif. SINE est verbe ou préposition;
   - soit régir tantôt l'accusatif, tantôt l'ablatif (SUPER, SUBTER, INSUPER), tantôt le génitif, tantôt l'ablatif (TENUS).

Dans les deux derniers types se trouvent des ambiguïtés sémantiques que l'on ne peut résoudre actuellement par ordinateur. Seuls les groupes dans lesquels la catégorie grammaticale est certaine peuvent faire l'objet d'un traitement automatique; ainsi, sur un total de quarante-quatre prépositions, il faut en éliminer d'emblée trente-deux pour n'en garder que douze.

On peut se demander s'il y a quelqu'intérêt à établir un programme spécial pour si peu de mots. Il y a deux raisons qui justifient des études semblables. Tout d'abord, il est utopique de chercher des lois

générales permettant de lever un grand nombre d'ambiguïtés. C'est au contraire par l'étude de cas particuliers portant sur un petit nombre de phénomènes morphologiques ou syntaxiques, que l'on confiera progressivement aux ordinateurs le choix d'une analyse correcte. Plutôt que de chercher à établir quelques lois générales, il faut tenter de définir une multitude de règles à portée restreinte.

La seconde raison est que parmi les douze prépositions non-amphibologiques figurent celles qui sont les plus fréquentes dans la langue.

En effet, AB, EX, DE et IN ont une fréquence d'apparition beaucoup plus élevée que SUPER ou ADVERSUS et même que ANTE ou POST.

Le type d'amphibologie que l'on aura à traiter pour un substantif régi par une préposition non-ambiguë, variera en fonction de quatre éléments, à savoir le cas régi par la préposition, la déclinaison, le genre et le nombre du substantif.

A) *Prépositions régissant uniquement l'accusatif*

1. Pour les substantifs neutres à l'accusatif singulier ou pluriel, l'ordinateur propose les trois analyses de nominatif, vocatif et accusatif.

   Exemple : AD COLLOQUIUM

analyse	AD	70300	
	COLLOQUIUM	12A00	(nominatif)
		12B00	(vocatif)
		12C00	(accusatif)

   Pour les neutres de la quatrième déclinaison au singulier, l'ordinateur propose en outre les analyses du datif et d'ablatif.

2. La même ambiguïté entre le nominatif, le vocatif et l'accusatif existe pour les substantifs pluriels, masculins et féminins de la troisième déclinaison.

3. A la quatrième déclinaison, la forme de l'accusatif pluriel est susceptible de recevoir six analyses morphologiques différentes puisqu'elle est identique aux nominatif, vocatif et génitif singuliers ainsi que aux nominatif et vocatif pluriels.

4. Enfin, pour l'accusatif pluriel de la cinquième déclinaison, on trouve cinq analyses possibles qui sont le nominatif et le vocatif singuliers, le nominatif, le vocatif et l'accusatif pluriels.

Pour les cas d'amphibologie que l'on vient de voir, il faudrait deux règles pour éliminer un maximum d'analyses erronées.

Lorsqu'un substantif suit une préposition régissant l'accusatif, il faut :

a) éliminer toutes les analyses autres que l'accusatif, si les interprétations concurrentes sont le nominatif ou le vocatif, singuliers ou pluriels.
Cette règle permet de traiter tous les substantifs neutres à l'exception des singuliers de la quatrième déclinaison, les substantifs masculins et féminins de la troisième déclinaison et tous les substantifs de la cinquième;

b) garder uniquement les analyses de génitif singulier et d'accusatif pluriel s'il s'agit d'un substantif masculin ou féminin de la quatrième déclinaison. En effet, dans ce cas, l'analyse du génitif ne peut être éliminée puisqu'on peut toujours être en présence d'un complément déterminatif.

Ces deux règles ne sont applicables que lorsqu'il n'y a pas, au niveau du mot qui suit la préposition, d'ambiguïtés sémantiques se combinant avec des ambiguïtés morphologiques. On verra dans la suite quelle solution il faut adopter dans le cas où une même forme peut se rattacher à plusieurs lemmes.

3) *Prépositions régissant uniquement l'ablatif*

1. Identité du datif et de l'ablatif. C'est l'ambiguïté la plus fréquente puisque, dans toutes les déclinaisons, le datif et l'ablatif pluriels sont semblables. Il en va de même du datif et de l'ablatif singuliers de la deuxième déclinaison, des ablatifs en -I de la troisième déclinaison.

2. A la première déclinaison, l'ablatif singulier est identique au nominatif et au vocatif singuliers.

3. Les substantifs masculins et féminins de la quatrième déclinaison ont parallèlement un datif en -VI, un datif en V qui est identique à l'ablatif. Quant aux substantifs neutres de cette déclinaison, ils ont au singulier les cinq formes identiques dont il a déjà été question.

Règle du choix de l'analyse : lorsqu'un substantif suit une préposition régissant l'ablatif, on peut éliminer toutes les analyses autres que celle de l'ablatif, quelle que soit la déclinaison de ce substantif.

Exemple :	AB	HOSTIBUS	
	AB	70600	
	HOSTIBUS	13N00	(datif)
		13Ø00	(ablatif)

C) *Prépositions régissant tantôt l'accusatif, tantôt l'ablatif*

Pour ce type de prépositions, l'amphibologie est double puisqu'elle se situe à la fois au niveau du mot régissant et du mot régi. Il faut donc tenter de donner à l'ordinateur des règles destinées à faire un double choix.

Compte tenu de la description des amphibologies qui se trouve dans les deux paragraphes précédents, il faut éliminer d'emblée les substantifs neutres singuliers de la quatrième déclinaison pour lesquels un choix est impossible en raison de l'identité de l'accusatif et de l'ablatif.

Pour le reste, on suivra les règles suivantes :

1. l'accusatif sera choisi pour la préposition et pour le substantif lorsque ce dernier peut être soit un nominatif, soit un vocatif, soit un accusatif singulier ou pluriel à l'exclusion de tout autre cas;
2. lorsqu'il s'agit d'un substantif de la quatrième déclinaison, on conservera les deux possibilités

d'analyse pour la préposition et les analyses de génitif singulier et d'accusatif pluriel pour le substantif;

3. l'ablatif sera retenu pour la préposition et pour le mot régi si celui-ci peut être notamment un ablatif et cela sans prendre en considération ni le nombre ni la déclinaison du substantif.

Les six règles que l'on vient de voir pour les trois types de prépositions sont applicables lorsqu'il n'y a que des amphibologies morphologiques. En effet, si une forme est susceptible d'appartenir à plusieurs lemmes, l'ordinateur n'a aucun critère pour effectuer un choix.

Il semble cependant qu'à ce point de vue également, il soit possible d'éliminer un certain nombre de fausses analyses. Plusieurs possibilités sont à prévoir.

A) Amphibologie de lemmes de substantifs

Une même forme peut appartenir à plusieurs substantifs homographes ou non. La règle générale à appliquer sera de garder la ou les analyses qui sont compatibles avec le ou les cas régis par les prépositions.

Exemple 1 :

AD		CASTRA	
AD		70300	
CASTRA	1	11A00	substantif de la 1ère déclinaison nominatif, vocatif et ablatif singuliers
	2	11B00	
	3	11F00	
	4	12J00	substantif de la 2ème déclinaison nominatif, vocatif et accusatif pluriels
	5	12K00	
	6	12L00	

La dernière analyse est la seule compatible avec la préposition AD; les autres peuvent être éliminées.

Exemple 2 :	E	CASTRIS	1	11N00	datif
			2	11Ø00	ablatif
			3	12N00	datif
			4	12Ø00	ablatif

Dans cet exemple, deux analyses doivent être retenues; on peut écarter les analyses de datifs.

Exemple 3 :	IN	CASTRIS	
		IN	70300
			70600
		CASTRA	11N00
			11Ø00
			12N00
			12Ø00

Cet exemple diffère du précédent par le fait que la préposition peut être suivie, soit de l'accusatif, soit de l'ablatif.

L'ordinateur retiendra la seconde solution pour la préposition et les solutions deux et quatre pour le substantif.

Exemple 4 :	IN	CASTRA		
		IN		70300
				70600
		CASTRA	1	11A00
			2	11B00
			3	11F00
			4	12J00
			5	12K00
			6	12L00

Dans l'expression IN CASTRA, on doit considérer que formellement IN peut être suivi de l'accusatif

(analyse 6) ou de l'ablatif (analyse 3); par conséquent, on conservera les deux analyses de IN et les solutions 3 et 6 pour CASTRA.

B) Amphibologie entre un lemme de substantif et un lemme d'une autre catégorie grammaticale

Pour la forme CASTRA, l'ordinateur donne en plus des analyses déjà rencontrées :

CASTRA verbe CASTRO 51B21 (impératif présent)

La règle du choix sera : lorsqu'une préposition est suivie immédiatement d'une forme qui peut être son régime, on éliminera toutes les analyses concurrentes si elles ne peuvent en aucun cas être considérées comme le régime de la préposition.

Les quelques règles qui précèdent sont destinées à montrer les possibilités dans le domaine d'une automatisation plus poussée de l'analyse du latin. Elles font apparaître la complexité des solutions pour le cas relativement simple de la préposition suivie de son régime. Il va de soi que cet exemple est très favorable puisqu'on y trouve une certaine stabilité dans l'ordre des mots, ce qui est loin d'être le cas pour l'ensemble de la phrase latine.

Une autre remarque s'impose. Le processus de choix de l'analyse correcte ne pourra donner d'emblée des résultats parfaits; il devra être amélioré progressivement, en tenant compte des observations que l'on fera en traitant un grand nombre de textes.

Aussi, sera-t-il nécessaire que, pendant un certain temps encore, le philologue contrôle le travail de l'ordinateur.

CHAPITRE VII

Programme d'analyse automatique de la morphologie du français

Le traitement automatique de la langue française peut s'effectuer selon un système analogue à celui que j'ai réalisé pour le latin.

Avant d'esquisser les grandes lignes de ce système, il faut signaler une différence importante entre le français et le latin : l'existence des accents, du tréma, de l'apostrophe, etc. Ces signes jouent un rôle capital dans la compréhension d'un texte puisqu'ils peuvent avoir, soit une valeur sémantique, soit une valeur morphologique. Ainsi, ce sont les accents qui différencient et identifient les mots *âpres* et *après*; de même, dans la conjugaison, l'accent est la seule différence entre le participe passé et certaines formes personnelles *(chanté, chante).* Dès lors, il va de soi que les accents et les signes spéciaux doivent être perforés et confiés à l'analyse de l'ordinateur.

Le processus de décomposition en radical et désinence utilisé en latin peut être appliqué, sans modification, à la langue française. Toutefois, la morphologie du français, relativement pauvre si on la compare à celle du latin, n'impose pas une table des désinences aussi développée, sauf pour les verbes : c'est pourquoi ce chapitre traitera de l'analyse morphologique automatique du verbe français.

Dans la description qui suit, j'ai utilisé comme ouvrage de référence, pour la conjugaison des verbes français, le *nouveau Bescherelle* (16) parce qu'il présente une description extrêmement *formelle* de la morphologie du verbe.

La complexité de la conjugaison française est telle que le *Bescherelle* distingue 82 types de verbes. Il est clair qu'en vue d'un traitement automatique, il faut procéder à des regroupements afin de limiter autant que possible le nombre de séries de désinences.

On peut réaliser, pour le français, un système analogue au système latin; il faut pour cela :

1) un ou plusieurs radicaux par verbe;

2) des désinences capables d'être jointes à un grand nombre de radicaux.

1.- *Les verbes du premier groupe (-ER)* (tableaux 6 à 18 dans *Bescherelle*)

Pour ces verbes, une seule série de désinences suffit. La difficulté qui résulte d'une modification du radical (changement d'accent - *céder* -, redoublement de la consonne finale - *appeler* - ou modification de celle-ci - *placer* -) est résolue par la création de deux ou trois unités lexicales pour chaque verbe qui présente une de ces particularités. Cette façon de procéder est identique à ce qui se fait en latin où l'on a trois radicaux différents pour chaque verbe.

RADICAUX

Tableau n°		Radicaux
6	AIMER	AIM
7	PLACER	PLAC PLAÇ
8	MANGER	MANG MANGE
9	PESER	PES PÈS
10	CÉDER	CÉD CÈD
11	JETER	JET JETT
12	MODELER	MODEL MODÈL
13	CRÉER	CRÉ
14	ASSIÉGER	ASSIÉG ASSIÈG ASSIÉGE
15	APPRÉCIER	APPRÉC
16	PAYER	PAY PAI
17	BROYER	BROY BROI
18	ENVOYER *	ENVOI ENVOY

* Sauf indic. futur simple et conditionnel présent.

DESINENCES.

Indicatif				Condi-tionnel	Impé-ratif	Subjonctif		Infi-nitif	Participe	
Présent	Imparfait	Passé simple	Futur simple	Présent	Présent	Présent	Imparfait	Présent	Présent	Passé
E (É)	AIS	AI	ERAI	ERAIS	E	E	ASSE	ER	ANT	É
ES	AIS	AS	ERAS	ERAIS	ONS	ES	ASSES			ÉE
E	AIT	A	ERA	ERAIT	EZ	E	ÂT			ÉS
ONS	IONS	ÂMES	ERONS	ERIONS		IONS	ASSIONS			ÉES
EZ	IEZ	ÂTES	EREZ	ERIEZ		IEZ	ASSIEZ			
ENT	AIENT	ÈRENT	ERONT	ERAIENT		ENT	ASSENT			

2.- *Les verbes du deuxième groupe (-IR, -ISSANT)* (tableaux 19 et 20)

Cette catégorie comprend les quelque trois cents verbes en -IR, -ISSANT (type FINIR) et le verbe HAÏR.

En ce qui concerne les verbes du type FINIR, il n'y a guère de difficulté : un seul radical suffit; pour le constituer, il suffit d'enlever à la forme de l'infinitif, la désinence -IR (FIN-). Les désinences seront normalement -IS, -IS, -IT, -ISSONS ...

Pour le verbe HAÏR, la présence du tréma empêche un traitement par les mêmes désinences que FINIR. Il semble que la solution qui s'impose pour HAÏR est de faire de chaque forme une unité lexicale et de renoncer au processus de décomposition.

3.- *Les verbes du troisième groupe* (tableaux 21 à 82)

Le *Bescherelle* distingue dans ce groupe quatre catégories :

1) le verbe *aller* (tableau 22);

2) les verbes en -IR qui ont le participe présent en -ANT (tableaux 23 à 37);

3) tous les verbes en -OIR (tableaux 38 à 52);

4) tous les verbes en -RE (tableaux 53 à 82).

Si on laisse de côté le verbe ALLER qui requiert un traitement particulier, on trouve dans les trois dernières catégories du ***Bescherelle*** des verbes aussi différents que VETIR, SAVOIR, VALOIR ou encore CONCLURE. Pour presque tous ces verbes, il faudra placer dans le lexique plusieurs radicaux et créer un système de codes relativement complexe. En outre, la plupart de ces verbes représentent à eux seuls un type de flexion. Il s'agit donc principalement de verbes irréguliers pour lesquels on doit enregistrer dans le lexique un certain nombre de formes sans les traiter par le processus de décomposition. Je prendrai comme exemple des difficultés que l'on rencontre dans le troisième groupe : les verbes qui forment le passé simple en -US et leur participe passé en -U (tableaux 33, 34, 38, 40, 41, 42, 43, 44, 47, 48, 63, 68, 69, 71, 74, 76, 77).

RADICAUX

Tableau n°		Radicaux
33	COURIR	COUR
34	MOURIR	MOUR MEUR
38	RECEVOIR	REÇOI REÇ RECEV REÇOIV
40	POURVOIR	POURV POURVOI POURVOY
41	SAVOIR	S SAI SAV SAU SACH

Tableau n°		Radicaux
42	DEVOIR	D DOI DEV DOIV
43	POUVOIR	POUV PEUV P POUR PUISS
44	MOUVOIR	MEU MOUV MEUV M
47	VALOIR	VAL VAUD VAILL
48	VOULOIR	VOUL VOUD VEUILL
63	PLAIRE	PLAI PLAIS PL
68	CROIRE	CR CROI CROY
69	BOIRE	BOIV BUV B BOI
71	CONCLURE	CONCL CONCLU
74	MOUDRE	MOUL MOUD
76	VIVRE	VI VIV VÉC
77	LIRE	L LIS LI

DESINENCES

Indicatif				Condi-tionnel	Impé-ratif	Subjonctif		Infi-nitif	Participe	
Présent	Imparfait	Passé simple	Futur simple	Présent	Présent	Présent	Imparfait	Présent	Présent	Passé
S	AIS	US	RAI	RAIS	E	E	USSE	IR	ANT	U
S	AIS	US	RAS	RAIS	ONS	ES	USSES	OIR		UE
T	AIT	UT	RA	RAIT	EZ	E	ÛT	RE		US
ONS	IONS	ÛMES	RONS	RIONS		IONS	USSIONS			UES
EZ	IEZ	ÛTES	REZ	RIEZ		IEZ	USSIEZ			
ENT	AIENT	URENT	RONT	RAIENT		ENT	USSENT			

Une seule série de désinences suffit pour ces verbes, mais certains problèmes subsistent.

1) *Formes non analysées*

Avec les informations dont il dispose, l'ordinateur ne pourra tout analyser. C'est ainsi que les formes PEUX, PEUT, VEUX, VEUT et MORT, par exemple, ne seront pas traitées. Ces formes sont difficiles à décomposer en radical et désinence; elles devront être placées dans le lexique avec leur analyse complète.

2) *Fausses analyses*

La nécessité d'avoir plusieurs radicaux pour un même verbe peut amener l'ordinateur à produire des analyses totalement fausses pour certaines formes qui existent dans la langue et qui, en apparence, pourraient être traitées comme des formes verbales. Prenons deux exemples.

La conjonction MAIS sera analysée par l'ordinateur comme l'indicatif de MOUVOIR (radical M-,

désinence -AIS).

SAVANT sera considéré comme le participe présent de SAVOIR (radical SAV-, désinence -ANT).

Pour éviter de telles erreurs, il faut créer un code qui empêche la décomposition de formes telles que MAIS ou SAVANT. Ce code existe déjà en latin (c'est lui qui empêche que DOS soit considéré comme l'accusatif pluriel de DEUS, radical D-, désinence -OS). Par conséquent, de telles erreurs sont faciles à éliminer. Le seul problème est d'établir la liste exhaustive des fausses analyses possibles. Cela se fera de façon empirique.

Par un système analogue à ce qui précède, on arrivera aisément à réduire à deux ou trois types de flexion l'ensemble des tableaux que le *Bescherelle* donne pour les verbes du troisième groupe. Il semble dès lors qu'il n'y ait aucune difficulté majeure à établir, pour le français, un système de lemmatisation et d'analyse morphologique identique à celui qui est utilisé pour le latin. Le système de traitement et les règles de constitution des lexiques qui ont été mis au point pour cette dernière langue sont applicables, sans grandes modifications, à la plupart des langues flexionnelles.

UNIVERSITE DE LIEGE
Laboratoire d'Analyse statistique
des Langues anciennes

J. DENOOZ

NOTES

(1) R. BUSA S.J., *Sancti Thomae Aquinatis Hymnorum Ritualium, Varia Specimina Concordantiarum,* Milan, BOCCA, 1951, p. 10.

(2) B. QUEMADA, *La mécanisation dans les recherches lexicologiques* dans *Cahiers de Lexicologie,* 1959, 1, p. 22.

(3) L. DELATTE, *Editorial* de la *Revue de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur,* 1965, 1, p. 3.

(4) A.J. GREIMAS, *Les problèmes de la description mécanographique,* dans *Cahiers de Lexicologie,* 1959, 1, p. 61.

(5) Cf. par exemple à ce sujet, L. DELATTE, Et. EVRARD, A. BODSON et S. GOVAERTS, *Sénèque, De Clementia,* Index verborum et Relevés statistiques, La Haye, MOUTON, 1968, pp. 127-134.

(6) Et. EVRARD, A. BODSON, *Le programme d'analyse automatique du latin,* dans *Revue de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur,* 1966, 2, p. 45.

(7) L. DELATTE, *Actes du Séminaire international sur le dictionnaire latin de machine,* dans *Revue de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur,* 1969, 2, p. 67.

(8) R. MICHEA, *Dictionnaire automatique et annotation des textes,* dans *Problèmes de la Traduction automatique,* Paris, KLINCKSIECK, 1968, p. 112.

(9) G. GOUGENHEIM, *Structure grammaticale et traduction automatique,* dans *Problèmes de la Traduction automatique,* Paris, KLINCKSIECK, 1968, p. 61.

(10) S. GOVAERTS, A. BODSON, *Codification d'un texte latin sur cartes mécanographiques IBM-80 colonnes,* dans *Revue de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur,* 1966, 1, pp. 1-48.

(11) L. HIRSCHBERG, *Le Choix d'une analyse morphologique dans le traitement automatique des langues,* dans *Problèmes de la Traduction automatique,* Paris, KLINCKSIECK, 1968, pp. 35-53.

(12) L. HIRSCHBERG, *Le Choix d'une analyse morphologique dans le traitement automatique des langues,* dans *Problèmes de la Traduction automatique,* Paris, KLINCKSIECK, 1968, pp. 42-43.

(13) Cf. par exemple à ce sujet, J. BLOIS et J. BUYDENS, *Règles catégoriques de levées d'ambiguïtés pour l'analyse syntaxique automatique du français,* dans *Problèmes de la Traduction automatique,* Paris, KLINCKSIECK, 1968, pp. 7-34;
Et. EVRARD, *Actes du Séminaire international sur le dictionnaire latin de machine,* dans *Revue de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur,* 1969, 2, pp. 67-68 et pp. 71-74;
Et. EVRARD, *La fréquence des phénomènes grammaticaux est-elle constante ?,* dans *Actes du 1er Colloque international de Linguistique appliquée,* Nancy, 1966, pp. 157-162.

(14) J. BLOIS et J. BUYDENS, *Règles catégoriques de levées d'ambiguïtés pour l'analyse syntaxique automatique du français,* dans *Problèmes de la Traduction automatique,* Paris, KLINCKSIECK, 1968, p. 12.

(15) Et. EVRARD, *Actes du Séminaire international sur le dictionnaire latin de machine,* dans *Revue de l'Organisation internationale pour l'Etude des Langues anciennes par Ordinateur,* 1969, 2, p. 74.

(16) Le *nouveau Bescherelle, L'art de conjuguer,* Paris, HATIER, 1966.