LES METHODES DU CENTRE DE TRAITEMENT ELECTRONIQUE DES DOCUMENTS

par P. TOMBEUR

Directeur du Centre de traitement électronique des documents de l'U.C.L.

CREATION ET OBJECTIFS

Le Centre de Traitement Electronique des Documents a été créé à l'Université Catholique de Louvain en 1968. L'objectif initial du Centre était le traitement automatique de documents médiévaux. Dès le début cependant, le Centre fut amené à considérer l'aide qu'il pouvait apporter à des chercheurs d'autres domaines. On ne pouvait étudier les problèmes d'automatisation en se limitant au seul domaine médiéval. Les procédés de travail élaborés sont en effet applicables à la plupart des recherches prenant leur point de départ dans un donné textuel. Le Centre de l'Université de Louvain fut ainsi mis au service de l'ensemble des chercheurs en sciences humaines.

L'objectif fondamental du Centre est l'automatisation de la plus grande part possible des travaux que comporte l'étude des documents. Dès à présent le Centre a constitué toute une bibliothèque de programmes d'ordinateur qui permettent notamment l'établissement automatique d'une série d'instruments de travail réalisables quelles que soient la langue et la nature de la documentation étudiée (index, plusieurs types de concordances, relevés statistiques divers). Ces instruments de travail sont établis tantôt à partir d'un simple enregistrement des données, tantôt à la suite d'analyses plus ou moins approfondies. Dans ce cas ils contiennent notamment les résultats d'analyses lexicographiques, morphologiques, syntaxiques ou stylistiques, de même que des classements et des analyses de données documentaires. Les recherches du Centre portent en outre sur les problèmes de comparaison de textes et de traditions manuscrites, sur les recherches de sources et, de façon générale, sur l'analyse statistique des documents étudiés.

Nos programmes sont actuellement rédigés en COBOL. Il va sans dire que l'évolution du matériel mis à notre disposition, les changements qui peuvent intervenir dans les systèmes d'exploitation, l'évolution des langages de programmation, de même qu'une critique constante des résultats obtenus nécessitent une adaptation et une amélioration continuelle des programmes existants. Les programmes d'ordinateur sont à la disposition des chercheurs. Le Centre collabore à la préparation et à l'exécution de leurs travaux.

Extrait de la Revue (R.E.L.O.)
VII, 1 à 4, 1971. C.I.P.L. - Université de Liège - Tous droits réservés.

EQUIPEMENT

Pour réaliser ces objectifs, le Centre dispose actuellement d'un ensemble de machines à cartes perforées et des ordinateurs IBM 360/40 et 360/44 du Centre de Calcul de l'Université. Nous utilisons principalement le 360/40 équipé pour le moment d'un lecteur et d'un perforateur de cartes, de cinq armoires à bandes magnétiques à neuf pistes, de quatre armoires à disques magnétiques et d'une imprimante.

METHODES D'ENTREE DES DONNEES

Les problèmes d'entrée des données ont une importance capitale. De la méthode d'entrée des données vont dépendre les possibilités de traitement ultérieures.

Que faut-il enregistrer ? Les divers types d'enregistrement

Avant de savoir comment on va enregistrer un document, se pose la question du document même qui est à enregistrer. En effet, celui qui désire étudier un document à l'aide de l'ordinateur peut envisager trois manières différentes de procéder :

- réaliser un enregistrement partiel de certains éléments d'un document;
- enregistrer un ensemble de données qui représentent les résultats d'une analyse documentaire de ce document;
- enregistrer l'ensemble du document.

Dans le premier cas, on ne retient que certains éléments contenus dans un document et ces seuls éléments seront enregistrés tels quels. Nous avons réalisé de la sorte des enregistrements de registres paroissiaux et civils : on n'a retenu que les seuls éléments qui intéressaient le chercheur voulant exploiter cette documentation. Quand il s'agit de textes, on peut enregistrer ainsi par exemple les seules citations; pour telle collection de droit canonique, on n'enregistrerait que les seules rubriques; dans un texte philosophique, on ne retiendrait que les seuls termes désignant les attributs de Dieu; dans tel corpus, on ne prendrait que les seuls *incipit* et *explicit*. De tels enregistrements peuvent évidemment être très utiles; ils offrent cependant le désavantage de ne permettre généralement que les seuls résultats visés directement par le chercheur. Le plus souvent ces enregistrements partiels ne pourront être exploités à des fins multiples.

Dans le second cas, si l'on désire enregistrer les résultats d'une analyse documentaire, il faut soumettre au préalable les documents à une telle analyse; celle-ci doit dégager les mots clés littéralement ou implicitement contenus dans un document. Ces mots clés ou descripteurs sont enregistrés avec les références aux documents qui les contiennent. On est bien obligé de se demander aussitôt s'il est possible de réaliser de la sorte des analyses exactes, complètes et basées sur des critères uniformes. C'est pourquoi l'on devine immédiatement que les possibilités d'application d'une telle méthode concernent avant tout des types

particuliers d'enregistrement et d'analyse, par exemple ceux qui sont à but essentiellement pratique. Et dans ces cas, l'on concevra même que les ordinateurs eux-mêmes vont dégager les mots clés contenus dans les documents étudiés. On perçoit aussi que cette méthode admet nécessairement *x* probabilités d'erreur, de confusion ou d'oubli. Par contre, quand il s'agit de l'étude d'un texte, d'une étude rigoureuse et minutieuse d'un document donné, il apparaîtra extrêmement difficile d'assurer l'analyse au moyen des seuls descripteurs, tout au moins si l'on désire dépasser le stade d'une interrogation partielle. Il faudra alors préférer se livrer à des analyses plus précises, et finalement plus simples, quitte à essayer de dégager ultérieurement les vrais mots clés d'un document, en entendant alors cette expression *stricto sensu,* c'est-à-dire les mots qui par leur fréquence exceptionnelle par rapport à une norme donnée, caractérisent ce document.

La troisième méthode de travail consiste, comme nous l'avons dit, dans l'enregistrement de l'ensemble d'un document. Chaque fois que cela est possible, il semble préférable d'éviter les inconvénients que nous avons cités et de procéder ainsi par enregistrement intégral : par exemple enregistrer l'ensemble des mots d'une oeuvre afin de pouvoir étudier cette oeuvre sous tous ses aspects. Telle est la pratique courante du Centre de Traitement Electronique des Documents, sans que nous excluions cependant ni l'enregistrement dit documentaire, ni l'enregistrement partiel. Deux éléments entraîneront en effet l'application de ces méthodes : le volume ou la nature d'une documentation donnée.

Comment enregistrer les documents ?

Ayant ainsi répondu à cette première question : que va-t-on enregistrer ? , il faut résoudre le problème du comment : comment va-t-on enregistrer les documents ? Quels supports doit-on choisir ? Comment ordonner les enregistrements sur ces supports ?

Les supports d'enregistrements les plus courants sont la carte perforée, la bande perforée, la bande magnétique et le disque magnétique. Un élément capital dans le choix d'un support d'enregistrement est le volume de la documentation à traiter. Plus ce volume est important, plus est contraignante la nécessité de disposer de supports utilisables dans des unités d'entrée et de sortie rapides. Le lecteur et le perforateur de cartes, le lecteur et le perforateur de bandes perforées sont des unités d'entrée-sortie lentes, tandis que les unités à bandes magnétiques et celles à disques sont des unités d'entrée-sortie rapides. Des documentations de grand volume ne pourront donc raisonnablement être traitées qu'avec des bandes magnétiques et des disques magnétiques.

Le problème que l'on rencontre immédiatement est celui de savoir comment on va pouvoir enregistrer ses documents sur de tels supports. Les enregistrements sur disques nécessitent au préalable l'existence d'un enregistrement sur un autre support. Les enregistrements sur bandes magnétiques peuvent actuellement être constitués de deux façons, soit à partir d'un enregistrement préalable sur un autre support, le plus souvent la carte perforée, soit directement à l'aide d'une encodeuse sur bande magnétique. Dans les deux cas, il s'agit d'un transcodage manuel. Seul le lecteur optique supprimerait la nécessité d'un tel

transcodage. Soulignons immédiatement qu'au stade actuel la lecture automatique de documents courants, tels notamment des textes, non transcrits préalablement soit avec des caractères spéciaux, soit avec une encre spéciale, n'est réalisée que de manière exceptionnelle. On constate qu'en outre une préédition des documents à enregistrer demeure indispensable dans la toute grande majorité des cas. De nombreuses données qui ne doivent pas être enregistrées accompagnent souvent le texte d'une page d'édition; sans un travail de préparation à l'enregistrement, on se trouverait encombré par quantité de fausses données. D'autre part la lecture optique, même dans son emploi actuel, constitue un procédé fort coûteux.

La première étape : la perforation des cartes

Jusqu'à présent le Centre de Traitement Electronique des Documents procède dans une première étape à la perforation de cartes. Quand il s'agit de textes, les quatre premières colonnes d'une carte sont réservées à un code d'identification; les colonnes 5 à 79 à la transcription du texte en continu; et de façon générale, et ce pour des raisons de facilité et de sécurité, la colonne 80 est réservée à la perforation de codes qui permettront la référenciation automatique de chaque mot du texte. Dans une phase de préenregistrement on aura déterminé le type de référenciation que l'on souhaite; on aura de même dû résoudre certains problèmes particuliers qui peuvent apparaître dans les éditions de textes. En procédant par perforation en continu, l'enregistrement d'un texte exige à la fois un minimum de temps de perforation et un minimum de cartes, ce qui limitera le temps nécessaire à la lecture des cartes en ordinateur.

Pour l'étude de documents particuliers nous procédons également dans un premier temps à la perforation de cartes; mais dans ces cas la nature de la documentation a déterminé préalablement les zones de perforation des données; une partie de carte, une ou plusieurs cartes sont affectées à l'encodage d'un enregistrement logique.

La seconde étape : la constitution de bandes magnétiques

La seconde étape du travail consiste pour nous dans le transfert du contenu de la carte perforée sur la bande magnétique. Ce transfert est réalisé par l'ordinateur. La carte perforée n'est donc pour nous qu'un intermédiaire entre le document d'origine et la bande magnétique.

Nous étudions pour le moment la possibilité d'enregistrer directement notre documentation sur une bande magnétique utilisable en ordinateur. Ces encodeuses existent maintenant depuis plusieurs années; les dernières machines lancées sur le marché semblent devoir retenir particulièrement notre attention. On enregistrerait ainsi directement un document sur bande magnétique en affectant à chaque enregistrement logique, par exemple à chaque mot, une zone aussi étendue qu'on le souhaite (on pourrait aller jusqu'à 240 positions par "record"). Le problème cependant n'est pas simple et il faut se garder soigneusement de se laisser abuser par la publicité ! Si on travaille avec une encodeuse, il faut se rendre compte que l'on ne peut envisager de commencer immédiatement le traitement, sous peine de mal travailler, c'est-à-dire de travailler lentement. Ces encodeuses opèrent en effet avec un facteur de groupage 1, alors qu'il faut

pouvoir déterminer en ordinateur le facteur de groupage optimal permettant de travailler le plus rapidement possible.

A l'occasion de cette remarque, je voudrais souligner l'évolution constante et proprement extraordinaire de toutes les techniques permettant le traitement automatique de l'information. Telle difficulté d'aujourd'hui se trouve résolue demain. Il importe donc avant tout de tracer une voie qui permette l'étude à la fois aussi polyvalente et aussi approfondie que possible : cette voie pourra toujours être suivie, quelles que soient les modifications de support, de machines, de programmes.

Avec la bande magnétique on ne ressent plus la contrainte des 80 colonnes affectées à chaque carte. On peut déterminer des zones aussi larges que nécessaires. Pour le traitement de textes ou de certains documents apparentés, nous réservons pour chaque "mot" une zone de 92 positions. Le programme d'ordinateur qui exécute le transfert de la carte à la bande réalise en même temps la référenciation de chaque mot, tout en réservant des zones vierges pour des analyses ultérieures. L'ordinateur référencie un document a la suite d'une carte d'initialisation, qui lui indique les références du premier enregistrement.
28 positions sont réservées à la référenciation : code d'oeuvre, livre, chapitre ou page, ligne ou vers, n° du mot dans la ligne ou le vers, n° dans la phrase, n° dans l'oeuvre; 2 positions sont réservées à des mots dits "oubliés"; 2 positions contiennent l'indication de la longueur du mot, que l'ordinateur calcule en même temps; 20 positions sont destinées au mot lui-même et toutes les positions restantes sont réservées à des analyses ultérieures. Il est en effet possible , à n'importe quel moment, d'enregistrer de nouvelles données sur une bande par forme de mise à jour. La mise à jour consiste à effacer, compléter ou modifier certains éléments enregistrés. Des programmes de mise à jour indiquent à l'ordinateur quelles sont les zones qu'il doit mettre à jour. Sans un programme de mise à jour, tout nouvel enregistrement sur une bande effacerait immédiatement les données préalablement enregistrées. Ici encore ce sont actuellement des cartes qui contiennent les adresses des mots à mettre à jour et les mises à jour proprement dites.

Etapes ultérieures

Un tel système de travail permet d'enrichir continuellement une bande-texte ou une bande-document initialement établie. La facilité et la rapidité des copies de bandes permettent d'envisager également des analyses toutes différentes pour un même donné. A cause de la rapidité de la lecture et de l'écriture sur bandes - 90.000 caractères à la seconde - tous nos traitements en ordinateur se font avec entrée et sortie sur bandes. Nous n'employons actuellement les disques magnétiques que pour les programmes et les tris. Les tris sur disques sont en effet extrêmement rapides. Les résultats des tris sont écrits sur bandes.

Tous nos *output,* toutes nos sorties sont de même inscrites sur bande; de l'unité à bande elles peuvent être transférées à l'imprimante et les résultats sont alors imprimés sur listings.

Ces méthodes d'entrée et de sortie permettent d'envisager le traitement de très grands ensembles. Elles permettent en outre d'enregistrer, de conserver, et d'analyser des documentations susceptibles d'intéres-

ser des chercheurs aux préoccupations diverses. Ceux-ci pourront, à la suite de leurs analyses, enrichir cette documentation préalablement établie.

TRAITEMENT DES DONNEES

Référenciation automatique

Le premier programme d'ordinateur auquel un document est soumis exécute actuellement à la fois le transfert de la carte perforée a la bande magnetique et la référenciation automatique de chaque élément (le plus souvent de chaque mot) d'un document. A la fin de ce passage, on dispose donc d'une bande magnétique où chaque élément du document est enregistré et référencié selon les normes que nous avons déja expliquées (1). Un programme d'impression permet d'obtenir les references de n'importe quelle partie du document ou de l'ensemble de celui-ci. Un listing d'ensemble pourra servir à la notation manuelle des analyses que l'on désirera éventuellement enregistrer sur la bande magnétique lors d'une prochaine étape.

Le chercheur qui n'envisage pas ces analyses ultérieures introduira à ce moment sur la bande les notations qui permettent de distinguer certains éléments de sa documentation qui doivent être caractérisés d'une maniere spéciale. Le cas le plus fréquent est la notation des citations et d'éléments de l'apparat critique. Des cartes de "mise à jour" porteuses des codes appropriés permettent l'introduction de ces notations. Celles-ci seront plus ou moins détaillées selon les cas. Ainsi pourra-t-on se contenter de noter d'un code identique toutes les citations ou, au contraire, distinguer chaque type de citations selon un code particulier (citations de l'Ancien Testament, citations du Nouveau Testament, citations de l'antiquité, citations d'auteurs contemporains, etc.).

Index, concordances, relevés statistiques divers

L'enregistrement réalisé permet d'établir automatiquement les divers instruments de travail réalisables sur la seule base des occurrences référenciées d'un document : divers types d'index, de concordances et de relevés de fréquences de formes (non groupées sous un lemme), ou des éléments contenus dans une forme, relevés des longueurs de formes et des longueurs de phrases, etc. La constitution de ces instruments de travail n'exige de la part du chercheur que le seul enregistrement sur bande magnétique. Or ces instruments sont déjà susceptibles de rendre de très grands services et de permettre des études nombreuses.

Le Centre de Traitement Electronique des Documents a établi les programmes nécessaires à l'élaboration de plusieurs de ces instruments de travail. Je ne peux ici les énumérer tous; je voudrais cependant attirer rapidement l'attention sur les programmes que nous avons constitués en matière de concordances. Nous avons établi plusieurs types de concordances, fournissant des contextes plus ou moins larges : x mots

avant la forme et *x* mots après la forme; toutes les phrases où la forme figure; la forme accompagnée d'un contexte "optimalisé". Chacune de ces concordances a son intérêt propre; pour l'usage général la première concordance présente le désavantage d'être trop peu élaborée, la seconde d'être trop vaste, d'où l'idée d'une concordance "optimalisée". L'ordinateur établit la concordance de chaque mot en partant de la phrase où il figure; il reconnaît les limites de chaque phrase grâce à l'enregistrement initial de la ponctuation forte et de la référenciation des mots selon leur numero d'ordre dans la phrase. Pour chaque mot il dispose de 100 positions : il fournira donc un contexte maximum de 100 signes. Tout mot contenu dans une phrase de moins de 100 signes figurera dès lors dans la concordance avec toute sa phrase; pour les phrases de plus de 100 signes, l'ordinateur calcule le contexte le plus large possible jusqu'à concurrence de 100 signes. Pour les mots qui sont en début de phrase, il donne le contexte qui suit; pour les mots qui sont en fin de phrase, il note le contexte qui précède; pour toutes les autres positions, dans des phrases de plus de 100 signes, il entoure le mot d'*x* mots avant, d'*y* mots après. Il faut noter que pour les mots qui figurent dans des phrases longues, le chercheur peut compléter le contexte d'un mot en se référant dans la concordance au contexte fourni pour un autre mot situé avant ou après lui; il peut dès lors restituer l'ensemble de la phrase par le recours à la seule concordance. Comme les autres instruments de travail, cette concordance peut être triée automatiquement selon des arguments divers; ainsi selon l'ordre alphabétique de la forme, selon l'ordre du texte, selon le contexte lui-même, ou selon tout autre critère préalablement fourni.

Analyse lexicographique

Ainsi que je l'ai déjà noté, les instruments de travail constitués de la sorte à partir de la bande-texte initiale ne comportent aucune analyse. Celle-ci peut être exécutée au cours d'une phase ultérieure et comporter plusieurs niveaux. Pour l'étude d'un texte, on sera le plus souvent obligé de se livrer au moins à une analyse lexicographique fondamentale. Celle-ci consiste dans la "lemmatisation" du texte, c'est-à-dire dans le fait d'attribuer à chaque occurrence la forme canonique du mot, telle qu'elle figure dans un dictionnaire de référence. L'étape de la lemmatisation n'inclut pas cependant la notation de toutes les distinctions qui peuvent être établies au niveau du lemme. A ce stade on ne peut apporter de solutions valables aux problèmes posés par la multiplicité de sens ou par l'évolution sémantique de certains mots. Ces problèmes de polysémie, distincts de ceux d'homographie, sont réservés à une étape d'analyse ultérieure. Il faut en effet disposer a cette fin d'une liste generale du vocabulaire d'une oeuvre ou d'une epoque.

Il sera souvent utile de faire figurer à côté du lemme la notation de la catégorie grammaticale à laquelle le mot appartient. Cette notation peut être limitée à l'identification suivante : mot significatif, mot plein, ou mot outil, mot fonctionnel. Dans certains cas on accompagne le lemme d'une analyse morphologique complète.

Jusqu'à présent le Centre a procédé par analyse manuelle transcrite sur cartes et transférée sur la bande-texte initiale, tout en travaillant à l'établissement de programmes d'analyse automatique. Le problème est extrêmement complexe, spécialement pour le latin; en latin en effet, l'homographie est particulière-

ment abondante et l'ordre des mots, le plus souvent, ne permet pas de reconnaître automatiquement une syntaxe, et donc de procéder à des analyses certaines. Rares sont les mots qui ne sont pas susceptibles formellement de faire l'objet de plusieurs analyses. Pour le moyen âge le nombre d'analyses formellement possibles est encore plus important. On sait que les textes médio-latins se caractérisent en outre par une très grande variété orthographique. Il faut dès lors veiller particulièrement aux homographies non relevées. Aussi le problème est-il de concevoir le système qui équilibre le plus heureusement la masse d'informations brutes que l'ordinateur peut nous donner (ainsi toutes les analyses formellement possibles), et l'ampleur de l'intervention requise de la part du chercheur. Notons en passant que ce type de problème se pose de façon tout à fait générale dans le domaine du traitement automatique de l'information.

Le système que nous avons conçu actuellement, doit permettre la lemmatisation automatique des formes non ambigues, la lemmatisation à titre provisoire des formes ambigues les plus courantes. Le chercheur dispose d'un listing qu'il complète et corrige. Ces ajoutes et ces corrections sont ensuite transférées sur la bande magnétique à partir de cartes comportant uniquement l'analyse et l'"adresse" du mot auquel l'analyse se rapporte. Cette adresse est constituée par le numéro du mot dans l'oeuvre.

Soulignons que d'autres types de lemmatisation sont possibles, même quand il s'agit d'un texte. On peut d'ailleurs, grace a des copies de bandes, executer pour un meme donne deux lemmatisations différentes, l'une correspondant au type traditionnel rattachant telle forme a tel lemme tel qu'il figure dans un dictionnaire de référence, l'autre rattachant par exemple tel mot à tel concept, à telle notion-titre. Dans une étude sur des traductions la lemmatisation peut notamment consister a ramener sous le mot propre à une langue les diverses traductions de ce mot dans une autre langue. On pourra de même procéder à une "lemmatisation" particulière quand il s'agit d'enregistrements non textuels, tel l'enregistrement d'une documentation de type archéologique. Le système général est en effet fort souple; il est susceptible d'être appliqué à des situations extrêmement diverses.

Quand le chercheur dispose d'une bande où le texte se trouve analysé du point de vue lexicographique, il peut constituer sur de nouvelles bases les divers types d'instruments de travail que j'ai évoqués : index, concordances, relevés de fréquence, etc. Chacune des analyses enregistrées peut être exploitée de façons diverses. La notation des catégories grammaticales ou de la seule distinction mot significatif-mot outil permet par exemple de constituer une concordance pour les seuls mots significatifs et un simple *index verborum* pour les mots outils.

Je voudrais m'attarder un instant aux documents que nous avons étudiés ou qui font l'objet de nos travaux actuels, et pour lesquels nous nous sommes limités à une étude lexicographique. Il s'agit principalement ici du travail réalisé pour le Comité National Belge du Dictionnaire du Latin Médiéval. Celui-ci fait en effet procéder à l'enregistrement de toutes les oeuvres médiolatines belges, dans le cadre d'un projet du Fonds de la Recherche Fondamentale Collective. Le 1er octobre 1969 nous avons commencé l'enregistrement des textes du IXe siècle. Dès à présent toutes les oeuvres du IXe et une grande partie des oeuvres du Xe siècle sont enregistrées sur bandes magnétiques et référenciées. Cela représente près

de 500.000 mots. Nous venons de terminer la lemmatisation et les concordances lemmatisées de toutes les oeuvres hagiographiques du IXe siècle.

Nous venons de commencer l'enregistrement du texte intégral du concile Vatican II. Les chercheurs disposeront au terme de ce travail d'une concordance complète de ce concile. Il en sera de même pour l'ensemble des conciles oecuméniques médiévaux dont nous avons commencé l'étude en collaboration avec M. Michel Mollat, professeur à la Sorbonne. Plusieurs textes isolés ont été soumis aux mêmes programmes, mais je ne puis m'y attarder ici.

Un type de lemmatisation particulière a été appliqué aux trois premiers livres du *De anima* d'Avicenne. Melle Van Riet, professeur à l'Université de Louvain, a réalisé un lexique latino-arabe. Ce lexique, dans son état actuel, livre d'une part les mots latins compris dans ce qu'elle a appelé leur "unité lexicologique" et d'autre part les racines ou les formes arabes correspondantes, notées dans leur translittération latine. Ce travail contribuera dans un premier temps à l'établissement de l'édition nouvelle des trois premiers livres du *De anima* (2).

Sur la base de tels enregistrements, quantité d'études sont possibles. Je voudrais, à titre d'exemple, évoquer l'intérêt d'une analyse des diverses couches du vocabulaire d'un texte, qui distinguerait par exemple dans tel texte philosophique le vocabulaire platonicien, aristotélicien, augustinien, etc. On peut également se livrer à des comparaisons automatiques d'index d'oeuvres diverses.

Analyses morphologiques, syntaxiques, stylistiques et autres

Le traitement automatique des textes ne s'arrête pas à la seule étude du vocabulaire; le chercheur peut, s'il le désire, se livrer à une analyse approfondie de son texte. Au point de vue de l'étude philologique, nous avons mis au point un système d'analyses morphologiques, syntaxiques et stylistiques approfondies. Les principes de codage et un certain nombre de codes utilisés au CETEDOC pour l'analyse d'un texte sont directement inspirés des travaux réalisés au Laboratoire d'Analyse statistique des Langues anciennes de l'Université de Liège. Les phénomènes morphologiques, syntaxiques ou stylistiques se répartissent selon un ordre de précision croissante. Le chercheur peut se limiter à tel ou tel type d'analyse, et pour chaque type, envisager des niveaux d'approfondissement divers.

Quand on veut disposer d'une concordance dans laquelle, sous leur lemme respectif, les formes sont classées selon l'ordre morphologique, il faut se livrer au préalable à une analyse morphologique. Celle-ci peut être complète (distinguer toutes les catégories grammaticales et toutes les sous-catégories), ou simplifiée (englober seulement les critères fondamentaux nécessaires au classement). Notons une fois de plus que le chercheur peut toujours ultérieurement enrichir sa documentation par forme de "mise à jour" : sur la même bande magnétique, un premier type d'analyse peut être automatiquement remplacé par un autre type d'analyse, plus complexe.

L'exemple que je désire vous présenter est celui de la concordance des *Auctoritates Aristotelis.* Melle Jacqueline Hamesse, du Centre De Wulf-Mansion de l'Université Catholique de Louvain, a enregistré l'édition qu'elle a réalisée d'un important florilège philosophique médiéval qui comporte quelque 3.000 citations et près de 30.000 mots. Ce sont les *Auctoritates Aristotelis, Senecae, Boethii, Platonis, Appuleii Africani, Porphyrii et Gilberti Porretani.* Une analyse lexicographique et morphologique lui a permis de constituer une concordance de tous les mots significatifs et un index des mots outils. Elle a de même enregistré sur bande magnétique les résultats de ses recherches concernant l'identification des citations. L'ordinateur a pu constituer dès lors des index d'identification : ceux-ci permettent au lecteur de la concordance de connaître l'auteur prétendu et l'auteur réel de chaque citation (et même la référence précise à l'auteur réel) et d'étudier, grâce aux mises en parallèle de l'auteur prétendu et de l'auteur réel, les fausses attributions médiévales. Cet ensemble de documents sera prochainement publié à partir des résultats imprimés par l'ordinateur, ce qui permettra la diffusion d'un instrument de travail particulièrement précieux pour l'identification des citations contenues dans les oeuvres philosophiques médiévales. D'autres florilèges seront étudiés selon les mêmes méthodes. De même plusieurs autres textes sont soumis aux mêmes programmes. Citons la *Legenda de origine ordinis fratrum servorum Virginis Mariae* (1317), le texte du procès de Jeanne d'Arc, le *De paenitentia* de saint Ambroise pour la collection des Sources Chrétiennes, etc.

L'étude syntaxique détermine la fonction de chaque mot. L'analyse syntaxique est particulièrement développée pour les verbes subordonnés. Le chercheur enregistre les codes syntaxiques sur la bande magnétique préalablement constituée par forme de "mise à jour". Il en va de même pour l'analyse stylistique. Celle-ci comprend généralement la notation de l'ordre des mots et de l'ordre des propositions. Cette notation est en relation directe avec celle de la fonction. On peut noter en outre des effets de style particuliers, tels que certains transferts de catégorie grammaticale, la présence de membres rimés, l'emploi de formes rares ou recherchées, etc.

Comme nous l'avons vu précédemment en ce qui concerne la lemmatisation, le chercheur pourrait concevoir l'analyse syntaxique d'une façon différente; le plus souvent il aurait avantage à mettre, ici également, deux formes d'analyse en parallèle. Le philosophe peut ainsi se livrer à une étude de la syntaxe du raisonnement; celle-ci viendrait compléter l'étude de la syntaxe de la langue. Sur la base, par exemple, d'un manuel de logique qui distinguerait les divers types de raisonnements comme autant de formes syntaxiques, il pourrait classer les diverses assertions d'un texte selon les formes de raisonnement logique dont elles relèvent.

Nous avons vu que l'automatisation de la lemmatisation et de l'analyse morphologique ne peut être que partielle. L'intervention du chercheur demeure de toute façon importante. Etant donné que l'analyse syntaxique et plus encore l'analyse stylistique exigent la compréhension des moindres finesses d'un texte, elles ne peuvent généralement faire l'objet que d'une automatisation très limitée. Le chercheur note les codes d'analyse; ceux-ci sont ensuite enregistrés sur la bande magnétique où le texte et éventuellement d'autres analyses préalables se trouvent enregistrés.

De nombreux programmes d'ordinateur permettent d'exploiter une documentation analysée. Les premiers d'entre eux trient les analyses selon divers types de classement. Des programmes d'impression permettent d'imprimer le résultat de ces tris et de constituer par conséquent des listings où le chercheur trouve de façon ordonnée les phénomènes analysés. Si les analyses se trouvent enregistrées sur une bande qui contient en outre, à la suite de l'application du programme de concordance "optimalisée", le contexte de chaque mot, le chercheur peut disposer de listings qui permettent de connaître chaque type d'analyse, les formes concernées, leur référence et leur contexte. Pour une étude syntaxique, par exemple, le chercheur peut donc constituer de la sorte une véritable concordance des fonctions. Dans la mesure où des analyses concernant l'apparat critique, les citations ou les titres ont été enregistrées, on peut classer séparément les diverses formes faisant l'objet de ces analyses. Enfin divers programmes de comptages rendent possible la constitution de tableaux statistiques de toutes les réalités analysées. A la suite de l'application de ces divers programmes, on dispose donc d'un ensemble d'instruments de travail qui permettent la vérification et la critique des analyses, la mise en question des critères d'analyse, la connaissance statistique et à la fois l'identification de chaque type d'analyse (on dispose en effet de listings qui permettent constamment de passer de la réalité statistique d'un phénomène au phénomène lui-même, situé dans son contexte).

Plusieurs textes ont déjà été soumis aux diverses analyses que j'ai évoquées. Il s'agit principalement d'oeuvres issues des milieux monastiques belges du moyen âge : la Chronique de Saint-Hubert (XIIe siècle), les *Gesta abbatum Trudonensium* de Gislebert de Saint-Trond (XIIe siècle), les *Vitae Ursmari* d'Anson et de Rathier de Lobbes (VIIIe et Xe siècles). Cette dernière étude s'inscrit dans le cadre d'un programme de recherches portant sur les remaniements d'oeuvres par l'analyse philologique comparée de la version primitive et de la version remaniée. Un des chercheurs du Centre vient d'entreprendre l'étude de Sigebert de Gembloux. Nous allons commencer cette année, avec Dom Froger, l'enregistrement de la *Regula Magistri.* D'autres projets encore sont en cours. Nous avons également établi la concordance d'un texte en français portant sur la réforme universitaire. Quant aux textes proprement philosophiques, nous avons entrepris avec le Père Stainier, de la Commission Léonine, l'étude philologique approfondie d'oeuvres de saint Thomas. Dans un premier temps cette étude porte sur le *De rationibus fidei* et le *De substantiis separatis.* Il s'agit de constituer de la sorte les premiers jalons d'une étude du latin philosophique et théologique du XIIIe siècle. En ce qui concerne encore saint Thomas, un autre travail concerne l'étude comparative de plusieurs traditions manuscrites. Nous avons ainsi enregistré l'ensemble des leçons variantes attestées pour un fragment du *Compendium Theologiae* de saint Thomas d'Aquin par une tradition de 75 manuscrits. L'étude comparative de plusieurs traditions manuscrites fait en effet partie des recherches du Centre. Signalons enfin que M. le Professeur Wenin et Melle Jacqueline Hamesse vont commencer, en collaboration avec le Centre, l'étude de l'oeuvre de saint Bonaventure.

Tous les programmes que j'ai rapidement évoqués et qui permettent l'étude des textes cités sont dès à présent opérationnels. Plusieurs autres sont actuellement en cours d'élaboration. Ils concernent notamment l'étude de la structure des phrases et des propositions, la comparaison automatique de textes et de traditions manuscrites.

Quant au système des enregistrements partiels, je signalerai que les méthodes du Centre ont permis également l'étude d'une documentation d'un tout autre type : Melle Maryvonne Houbion a enregistré les données anthroponymiques du registre paroissial et du registre de l'état civil de Malonne au XIXe siècle. Grâce à l'analyse effectuée, elle a constitué un index de tous les prénoms attestés à Malonne à cette époque. Cet index note pour chaque prénom tous les éléments d'analyse retenus (correspondances entre les deux registres, origine du prénom, etc) et fournit un ensemble de fréquences.

CONCLUSION

Les possibilités offertes par l'emploi de l'ordinateur dans nos domaines de recherches sont extrêmement diverses. J'ai voulu vous expliquer brièvement nos méthodes. Cependant je ne peux m'empêcher de conclure en évoquant encore un autre point de méthode ... la joie que nous procure un tel travail. Au point de départ l'emploi de l'ordinateur nous oblige à des analyses rigoureuses; il nous force à une remise en question continuelle et de nos critères d'analyse et des analyses elles-mêmes. Il nous amène même à poser des questions fondamentales; il nous permet enfin de disposer d'instruments de travail exhaustifs qui rendent possible l'analyse approfondie des moindres caractéristiques d'un texte dans des limites de temps jadis insoupçonnables. Et pourtant tout ce domaine, aujourd'hui encore, est neuf; il se trouve par ailleurs confronté à un progrès technologique extraordinaire.

L'ordinateur est devenu pour nous un instrument de recherche indispensable. Recherche passionnante et humble. Rien ne serait d'ailleurs plus éloigné d'un véritable esprit de recherche qu'un sentiment de vaine suffisance. J'espère que les résultats d'aujourd'hui susciteront des applications et des adaptations multiples et qu'ils ouvriront la voie à des possibilités nouvelles.

NOTES

(1) Cf. *supra,* p. 5.

(2) Melle Van Riet a fait sur ce sujet une communication au Ve Congrès international d'Etudes arabes et islamiques tenu à Bruxelles en septembre 1970. Cette communication vient d'être publiée dans la *Revue philosophique de Louvain,* t. 68 (1970), pp. 521-535, sous le titre *Traductions arabo-latines et Informatique.*

SUMMARY

1. Establishment in 1968 of a Centre for the Data Processing of Documents at the Catholic University of Louvain; this Centre is at the disposal of all the human sciences. Up to the present the Centre has established a series of computer programmes, drawn up at the moment in COBOL, which permit the automatic processing of documents.

2. Equipment : IBM 360/40 and 360/44 (128 K) with a reader-puncher of cards, units of magnetic tapes and magnetic disks, a rapid printer.

3. Method of recording the data : many methods are possible (complete recording, partial recording, content-analysis). In most cases the Centre makes a complete recording of the document. The text is punched on cards without break (i.e. a many words per card), and then transferred on to a magnetic tape, where each word constitutes a distinct record. Each record is made up of many zones : a zone for the word itself, a zone for its references, and a zone for the analyses to be added later.

4. Method of processing of the data : automatic referencing, drawing up of an index, drawing up of differents types of concordances (especially the type of concordance in which the best context for each word is chosen by the computer), and the making of different statistical lists and tables. Recording of lexicographical, morphological, syntaxical and stylistic and other types of analyses. Classification and analysis of documentary data.

5. Enumeration of the principal documents studies.