CENTRE DE TRAITEMENT ÉLECTRONIQUE DES DOCUMENTS (CETEDOC). RECHERCHES EN COURS

PAUL TOMBEUR

La perspective fondamentale du Cetedoc est de développer au maximum le recours à l'informatique comme science auxiliaire des sciences humaines. Notre Centre doit donc faire face aux problèmes de traitement de l'information qui se présentent en linguistique et en philologie, en histoire et en sociologie, en philosophie et en théologie, en archéologie, etc. Notre rôle principal est dès lors de parcourir l'ensemble des étapes depuis l'énoncé d'un problème jusqu'aux exécutions, en passant par l'analyse formelle et technique, la programmation et les tests. Il s'agit de trouver les solutions informatiques pour les problèmes scientifiques qui sont de notre ressort.

Depuis le premier janvier 1974 (cf. les Actes du colloque de 1974, p. 153 sv.), nous avons ainsi été amenés à écrire des programmes nouveaux, à modifier ou à rendre davantage opérationnels des programmes anciens. Ces programmes ont été appliqués à quantité de documents de natures, de langues et d'époques diverses. Nous tenons à jour un fichier sur bande magnétique qui fait état de notre banque de données.

La programmation la plus neuve concerne le classement des manuscrits. Au-delà des réalisations notées il y a trois ans, nous

disposons maintenant d'une programmation dégageant automatiquement une généalogie de manuscrits. Dans les mois qui viennent, nous comptons expérimenter ces programmes sur une tradition manuscrite créée de toutes pièces.

Un point important à noter: nous avons, au cours de ces trois années écoulées, mis résolument l'accent sur l'informatisation des scientifiques. Un nombre sans cesse accru d'étudiants de licence et de chercheurs se sont, d'autre part, initiés à l'informatique.

L'élaboration d'une méthodologie informatique est pour nous chose capitale. Je me permettrais dès lors d'insister sur cette méthodologie en rapport avec la question posée sur *ordo*, en considérant la présentation qui suit comme un résultat de notre travail de ces dernières années. Avant de parler de résultats, il faut insister sur la méthode. La méthode informatique doit être élaborée en vertu des exigences scientifiques. Que l'on prenne garde à ne pas mettre entre parenthèses les exigences de la science auxiliaire en tant que telle. On pourrait faire ici un utile rapprochement avec la paléographie. Celle-ci n'a vraiment progressé, ne s'est même radicalement renouvelée qu'à partir du moment où elle a vraiment reçu ses lettres de créance. Trop souvent les philosophes, les théologiens, les historiens, voire même les linguistes réagissent vis-à-vis de la paléographie comme vis-à-vis de quelque auxiliaire auquel on ne prêterait en lui-même aucune attention particulière, et du coup méconnaissent à la fois cette discipline et les avantages réels qu'ils peuvent en attendre. Il en va de même pour l'informatique. Il faut se rendre compte que c'est la démarche scientifique même qui est en cause.

Voulant interroger un corpus textuel donné, nous suivons un ordinogramme précis:

1. L'enregistrement des formes, de toutes les formes, c'est-à-dire de tous les mots contenus dans le corpus (dans le cas, évi-

demment, où le corpus ne se trouve pas déjà sur un support traitable en ordinateur).

2. L'analyse lexicographique fondamentale, c'est-à-dire la lemmatisation, qui distingue les vocables mais non les diverses significations d'un même vocable. Cette analyse est accompagnée d'une analyse morphologique plus ou moins détaillée selon les cas.

A partir de ce stade, nous appliquons:

a) le programme de calcul de fréquences, permettant ultérieurement des études statistiques,

b) le programme de concordance «optimalisée», situant chaque formè dans un contexte optimalisé au sein de la phrase dont elle fait partie,

c) le programme de concordance «x-x», classant toutes les expressions identiques (ce qui peut déboucher notamment sur des analyses syntaxiques et stylistiques); à partir de là, un autre programme opère la sélection de toutes les expressions répétées. D'où l'étude, sous des aspects multiples, des co-occurrences (notamment la distribution des co-occurrences par éloignement progressif),

d) le programme de concordance-phrase opérant sur la base d'une interrogation (ex. toutes les phrases contenant le lemme *ordo*).

3. La description automatique du lemmaire. Nous entendons par lemmaire, l'ensemble des lemmes d'un corpus. Nous avons présenté au congrès de l'Association for literay and linguistic computing, qui s'est tenu à Oxford en avril 1976, un rapport intitulé «Elaboration d'une description automatique complète du vocabulaire latin». Le but de cette opération consiste, en faisant l'analyse et par conséquent l'explicitation du lemme proprement dit, à pouvoir disposer automatiquement de toutes les connotations d'un lemme donné - étymologie, première attestation, radical, synonymes, etc. -, et à interroger dès lors le vocabulaire d'un

corpus selon des critères nouveaux (un ex.: une interrogation pour *ordo* de tous les lemmes de même radical).

L'analyse formelle de cette description du lemmaire est terminée. Des premières applications partielles ont été réalisées pour le latin (ex. chronologie des lemmes). Il s'agit là évidemment d'un gros travail dont l'achèvement nous prendra encore beaucoup de temps. Celui-ci exige, en effet, la consultation des travaux lexicographiques qui sont à notre disposition.

4. L'analyse sémantique. C'est seulement à ce stade, après avoir parcouru toutes les étapes précitées et multiplié de la sorte les points d'observation, que l'on peut se livrer à une réflexion sur le sens; du même coup, l'on sort de l'ordinogramme informatique proprement dit. L'ordinateur, machine à traiter l'information, manipule celle-ci de multiples façons; il est en quelque sorte un instrument de syntaxe, non de sémantique. Tout sens fourni par la machine est un sens préformé. La seule chose que l'on puisse faire, c'est lui fournir les résultats de nos interprétations et soumettre celles-ci à de nouvelles manipulations.

Pour ce qui concerne le travail relatif à *ordo*, nous avons parcouru les étapes suivantes:

1. Examen et critique des rubriques consacrées à *ordo* dans les divers dictionnaires latins et répertoires de nature lexicographique (Ernout-Meillet, *Thesaurus linguae latinae,* Forcellini, *Oxford Latin Dictionary*, Gaffiot, Blaise, concordances publiées de la Bible, de Boèce, etc, Niermeyer, *Mittellateinisches Wörterbuch* et les dictionnaires nationaux concernant le moyen âge latin).

2. Elaboration d'une hypothèse de travail. L'idée fondamentale d'ordre est ce qui représente un rang. L'ordre, c'est ce qui est relié, en connexion, organisé.

3. Interrogation informatique d'un vaste ensemble de textes latins. Nous avons constitué ainsi un fichier de toutes les phrases qui contiennent *ordo*.[1]

Parallèlement à ces travaux, nous avons organisé des séminaires sur les thèmes suivants: la recherche de méthodes nouvelles pour la comparaison de textes de Thomas et de Bonaventure, l'analyse du vocabulaire, et enfin la possibilité de réaliser un dictionnaire de référence unique pour la langue latine (et ce sous l'impulsion directe du Lessico Intellettuale Europeo).

Depuis janvier 1974, nous avons publié plusieurs volumes dans notre collection «Informatique et étude de textes»:

— pour le corpus des conciles oecuméniques:

Michel Mollat et Paul Tombeur, avec le concours de Georges Mailleux et Christine Pellistrandi, *Les conciles Latran I à Latran IV. Concordance, Index verborum, Listes de fréquences, Tables comparatives*, 1974, XIX-225 p.

Id., avec le concours de Georges Mailleux, Hubert Maraite et Christine Pellistrandi, *Les conciles Lyon I et Lyon II. Concordance,...*, 1974, X-166 p.

Ph. Delhaye, M. Guéret, P. Tombeur, *Concilium Vaticanum II. Concordance,...*, 1974, XX-978 p.

— le tome II du Thesaurus Bonaventurianus:

Jacqueline Hamesse, *Breviloquium. Concordance, Indices*, 1975, X-431 p.

— en qui concerne les florilèges médiévaux:

Jacqueline Hamesse, *Auctoritates Aristotelis, Senecae, Boethii, Platonis, Apulei et quorundam aliorum*, tome II. *Index verborum*,

[1] Les attestations figurant dans les textes philosophiques ont été étudiées par Jacqueline Hamesse. Cfr. son rapport sur *Le concept ordo dans quelques oeuvres de saint Bonaventure* (cfr. *supra*, pp. 27 sqq.).

Listes de fréquences, Tables d'identification, 1974, XIII-137 p.

— pour le Corpus des Sources franciscaines:

le tome I, Georges Mailleux, *Thesaurus Celanensis, Vita prima, Legenda ad usum chori, Vita secunda, Tractatus de miraculis, Legenda sanctae Clarae virginis. Concordance, Index verborum, Listes de fréquences Tables comparatives,* 1974, XX-889 p.

le tome II, Jean-François Godet, *Sancti Bonaventurae Legendae maior et minor s.Francisci. Concordance,...* 1975, XV-452 p.

le tome III, Jean-François Godet et Georges Mailleux, *Legenda trium sociorum. Anonymus Perusinus, Fr. Juliani de Spira Vita s. Francisci, Sacrum commercium s. Francisci cum domina paupertate. Concordance,...,* 1976, X-395 p.

le tome V. Id., *Opuscula s. Francisci, Scripta s. Clarae. Concordance,...,* 1976, VIII-435 p.

— la Chronique de Saint-Hubert:

Paul Tombeur, *Chronique de Saint-Hubert. Concordance, Index verborum, Relevés statistiques,* 1974, XXII-518 p.

— enfin, l'étude consacrée à l'analyse littéraire d'un roman néerlandais, qui se termine par une intéressante postface de la romancière qui a accepté de réfléchir longuement sur son oeuvre passée à partir des résultats d'ordinateur:

Marc Geerinck, *Hella S. Haasses ontmoeting met de computer. Poging tot literairkritische analyse van De Verborgen Bron op grond van de automatische behandeling van de tekst. Met een nawoord door Hella S. Haasse,* 1976, 227 p.

Parmi les documents encore inédits, je signalerai la thèse monumentale de Jean Schumacher, *L'oeuvre de Sigebert de Gembloux. Etudes philologiques,* Louvain, 1976, 963 p., complétée par une très vaste documentation annexe de première importance présentant de manières diverses le vocabulaire de l'ensemble des oeuvres de Sigebert.

J'avais le plaisir de présenter il y a trois ans le premier volume de notre Nouveau répertoire des sources médiolatines belges. Le tome II a paru en 1976:

Index scriptorum operumque latino-belgicorum medii aevi, publié sous la direction de Léopold Genicot et Paul Tombeur, deuxième partie: *XI^e siècle*, par Paul Fransen et Hubert Maraite, Académie Royale de Belgique, Bruxelles, 1976, 279 p.

Dans les prochains mois paraitra la première partie du tome concernant les oeuvres du XII^e siècle.

La masse des textes soumis au traitement informatique ne fait que croître de jour en jour. Les domaines linguistiques sont de plus en plus divers. Nous espérons d'ailleurs diffuser prochainement la liste des oeuvres traitées.

Christian Wenin a présenté, dans son rapport, des travaux informatiques réalisés par l'Institut Supérieur de Philosophie en collaboration avec le Cetedoc (Aristote, Avicenne, Bonaventure). Je signalerai en outre que nous avons commencé l'étude du Pseudo-Denys, ainsi que celle de l'*Ethique* de Spinoza. Dans notre banque de données figurent également les commentaires grecs de l'*Ethique à Nicomaque* dans la traduction latine de Robert Grosseteste. Un Corpus particulièrement important dans le domaine de l'histoire des religions, est celui des pseudépigraphes grecs de l'Ancien Testament dont nous faisons actuellement le traitement systématique.

Un mot enfin de notre plus vaste corpus, celui de tous les textes médiolatins belges, c'est-à-dire essentiellement de tous les textes latins écrits en Belgique ou par des auteurs d'origine belge au cours de moyen âge. Nous avons terminé l'enregistrement de tous les textes narratifs des origines à 1200 et nous avons commencé l'étude et l'enregistrement des textes diplomatiques pour cette mê-

me période. Nous avons entamé l'élaboration du lemmaire général, c'est-à-dire la liste complète du vocabulaire attesté dans nos textes du VII[e] siècle jusqu'à la fin du XII[e]. Les inventaires linguistiques qui en résultent seront publiés au cours des années à venir. Ils préluderont à l'analyse sémantique, c'est-à-dire à l'élaboration de notre dictionnaire national.