RECHERCHES EN COURS AU « CENTRE DE TRAITEMENT ÉLECTRONIQUE DES DOCUMENTS » (CETEDOC)

par PAUL TOMBEUR et JACQUELINE HAMESSE

Je voudrais d'abord rapidement présenter le CENTRE DE TRAITEMENT ELECTRONIQUE DES DOCUMENTS de l'Université Catholique de Louvain, qui a été fondé en 1968, sur la base de l'expérience acquise au LABORATOIRE D'ANALYSE STATISTIQUE DES LANGUES ANCIENNES de l'Université de Liège, où j'ai pu travailler de 1961 à 1968.

Quelle est la perspective essentielle de notre Centre? C'est de concevoir l'informatique comme une science auxiliaire de la philologie, de la linguistique, de la philosophie, de la théologie. Il nous faut donc affiner cette science auxiliaire, la rendre aussi opérationnelle que possible. Aussi notre tâche principale consiste-t-elle à concevoir et à créer les programmes d'ordinateur qui permettent d'automatiser au maximum l'étude des textes. Ce faisant, nous devons veiller à faciliter le plus possible les problèmes d'input et d'output des données.

Nous travaillons avec les ordinateurs du Centre de Calcul de l'Université, c'est-à-dire actuellement encore un 360/40 IBM et un 370/155, qui vient d'être remplacé il y a deux semaines par un 370/158. De plus, nous avons sur place un ordinateur Philips de 72K, qui permet essentiellement de gérer les inputs, notamment à l'aide de terminaux à écrans cathodiques. Ce matériel n'est pas encore relié à l'heure actuelle avec le Centre

de Calcul, mais il servira dans quelque temps de terminal lourd pour le 370/158.

Voici très brièvement l'organigramme de notre système. Précisons que le cas, pour nous, le plus courant, est l'étude de textes de langues et de natures diverses. Nous avons de nombreux chantiers en cours; les chantiers se diversifient essentiellement d'après les « clients », qui s'adressent à nous, selon les problèmes posés par les chercheurs, problèmes qui nécessitent pour eux le traitement de telle ou telle documentation.

Je crois pouvoir passer très rapidement sur les diverses étapes de traitement d'un texte ou d'un ensemble de textes. Elles sont généralement bien connues. Après le choix du document d'origine, nous en faisons l'enregistrement intégral; dans la plus grande partie des cas, nous faisons actuellement l'encodage, grâce au système à écran cathodique, directement sur disque magnétique; le contenu du disque est ensuite transféré sur bande magnétique. Nous enregistrons les textes « en continu », les mots les uns à la suite des autres, en ajoutant, chaque fois qu'il y a lieu, les codes qui permettront ultérieurement une référenciation automatique de chaque mot.

Après la vérification et la correction du fichier enregistré, nous procédons à la segmentation du texte: l'unité d'enregistrement est désormais non plus un ensemble de mots, mais un seul mot. Celui-ci est référencié par l'ordinateur de façon multiple (ainsi selon l'oeuvre, le livre, le chapitre, la ligne ou le vers, le numéro dans la ligne ou le vers, le numéro dans la phrase et dans l'oeuvre); l'ordinateur calcule en même temps la longueur de chaque mot.

Quand le texte est ainsi référencié sur bande magnétique, nous introduisons, s'il y a lieu, les analyses qui caractérisent le texte de façon générale: les indications d'apparat critique, les résultats de l'étude des sources ou toute autre caractérisation distinguant des parties spécifiques.

Nous constituons ensuite une série d'instruments de trava

qui exploitent les diverses informations relevées (concordances diverses, index, relevés statistiques divers). Ces informations sont encore « brutes », les seules analyses effectuées sont de niveau général. A partir de ce stade interviennent dès lors les diverses phases d'analyse. Celles-ci se situent généralement dans la ligne traditionnelle et comportent des niveaux de complexité croissante: analyse lexicographique d'abord, morphologique, syntaxique, stylistique ensuite.

Je ne m'attarderai pas à la lemmatisation; le problème est connu de tous. Celle lemmatisation est, bien entendu, comme partout ailleurs dans la plupart des cas, automatique partiellement, étant donné que dans tous les cas, les lemmes proposés par l'ordinateur doivent être minutieusement vérifiés par les chercheurs pour voir si les analyses proposées sont correctes. Les dictionnaires automatiques qui permettent ces lemmatisations, sont enrichis au fur et à mesure des textes traités.

En ce qui concerne la lemmatisation, je soulignerai qu'à partir d'une première lemmatisation, que l'on peut considérer comme une analyse lexicographique fondamentale, nous envisageons, et cela a déjà été fait dans le domaine de certains textes anglais et français, une lemmatisation de type conceptuel ou thématique. Un type n'en supprime cependant pas un autre; l'ensemble des distinctions faites sont toujours retenues et par conséquent tous les critères utilisés peuvent toujours être critiqués.

Quant aux instruments de travail élaborés, j'insisterai particulièrement sur leur diversité. Ainsi avons-nous constitué notamment plusieurs types de programmes de concordances. Un de ces programmes est appliqué quasi systématiquement aux textes enregistrés. Sa caractéristique essentielle est de fournir pour chaque mot un contexte situé et calculé par l'ordinateur à l'intérieur de la phrase où le mot figure, ce qui permet au chercheur, par la seule consultation de la concordance, de retrouver la phrase entière où le mot figure; chaque fois que la phrase

entière ne figure pas comme contexte sous un lemme donné, parce que cette phrase est trop longue, il suffit de consulter la concordance à un mot situé au début ou en fin de phrase, pour retrouver le contexte-phrase complet.

En dehors des concordances-phrases proprement dites, nous utilisons beaucoup pour certaines recherches, la concordance que nous appelons « x-x », parce qu'elle fournit x formes avant le mot, x formes après le mot, tout en rapportant également la ponctuation. L'intérêt de cette dernière concordance est de permettre directement une étude des expressions. En effet, comme d'ailleurs pour tous nos programmes, cette concordance est un élément qui se situe dans une « chaîne de programmes »; ainsi, à la suite de l'application de ce programme « x-x », figurent notamment d'autres programmes qui permettent d'obtenir les expressions qui apparaissent plus d'une fois dans un même texte. Que signifie ici le mot expression? Le chercheur détermine lui-même ce qu'il entend par là, le nombre de mots qu'une expression doit comporter, et il peut évidemment faire varier les types de tris qu'il désire en obtenir.

Le degré d'automatisation des analyses morphologiques, syntaxiques et stylistiques est très variable selon le type d'analyse retenu. C'est évidemment dans le domaine de l'analyse morphologique que l'automatisation est la plus grande. Plus on approfondit l'analyse syntaxique, plus on s'aperçoit à quel point toute automatisation est difficile. On procède alors souvent par analyses manuelles. Une fois celles-ci insérées dans les fichiers, l'ordinateur peut en assurer une exploitation multiple, et cela demeure un avantage incomparable. Un des intérêts majeurs de cette exploitation automatique des analyses manuelles, est de nous amener à critiquer constamment les analyses qui ont été effectuées et en tout premier lieu le système même de codage qui a été appliqué. Le problème du codage des analyses syntaxiques est particulièrement complexe. Selon quel système faut-il coder les analyses syntaxiques? Je crois que, finalement, il est im-

portant de souligner, que dans une perspective générale d'automatisation, le problème du système linguistique de base suivi n'est pas essentiel, étant donné que de nombreux programmes de transcription, de transposition, de transcodage permettent de passer d'un type d'analyse à un autre.

Les exploitations des analyses sont très multiples. On peut obtenir de nouveaux types d'index, ou de concordances, ainsi notamment des concordances syntaxiques; au lieu de classer les contextes selon le lemme ou la forme, on peut les classer selon le code syntaxique, regroupant par exemple tous les compléments déterminatifs, les compléments d'objet direct, etc. Toutes ces informations sont groupées selon l'ordre désiré par les chercheurs. On se situe désormais dans un système où tout devient complémentaire, où une façon de procéder ne se trouve plus opposée à une autre.

Dans les perspectives mêmes de ce colloque, il est particulièrement important de souligner les divers types de regroupements des informations recueillies. Regroupements des informations que l'on pourrait appeler individuelles, en des « thesauri » généraux. A partir de ces « thesauri » constitués pour tel ou tel texte, on constitue un « thesaurus » pour l'ensemble des oeuvres d'un auteur, pour l'ensemble des textes de même époque, pour des textes de même genre littéraire, ou pour la production générale d'un époque ou pour des documents rédigés en une langue donnée, comme, par exemple, notre projet de constitution d'un dictionnaire du latin médiéval, auquel je ferai allusion de façon un peu plus détaillée dans quelques instants. Cette constitution notamment de lexiques groupés permet évidemment l'élaboration de dictionnaires de plus en plus étendus dans des domaines variés.

A chaque stade, nous avons également prévu des programmes de publication éventuelle; ainsi tous les livres de notre collection « Informatique et étude de textes » sont publiés directement à partir des listings imprimés par l'ordinateur.

Un mot, simplement, des autres systèmes que nous essayons d'automatiser en dehors du traitement des textes: je dirai uniquement qu'une de nos préoccupations est le problème de l'étude des traditions manuscrites; toute une série de programmes ont été constitués pour aider au classement des variantes et visualiser les types de regroupements des divers témoins d'un texte.

Parlons maintenant des corpus étudiés. A côté de toute une série de textes de langues diverses qui vont de l'antiquité jusqu'à nos jours, nous traitons systématiquement un certain nombre de grands ensembles: soit actuellement le corpus des textes philosophiques, dont vous parlera Jacqueline Hamesse, le corpus des sources franciscaines du XIII^{e} siècle, les conciles oecuméniques du XII^{e} siècle jusqu'à Vatican II, et enfin, le corpus du dictionnaire du latin médiéval. Je voudrais évoquer rapidement les étapes que nous suivons pour ce dictionnaire qui est réalisé dans le cadre et grâce à l'appui de notre Fonds National de la Recherche Scientifique et avec l'aide également de l'Académie Royale de Belgique. En Belgique, comme dans bien d'autres pays, il existe un Comité National du Dictionnaire du Latin Médiéval; c'est ce Comité qui a engagé l'ensemble de ce travail afin d'élaborer un dictionnaire du latin médiéval belge.

La première opération consiste à élaborer un répertoire de tous les textes à retenir. Il fallait, en effet, constituer un nouveau répertoire, qui se présente comme un « status quaestionis » des textes médiolatins. Je vous présente ici le premier volume qui vient de sortir de presse; il est consacré aux oeuvres du VII^{e} au X^{e} siècle. Nous espérons publier l'année prochaine le volume consacré aux oeuvres du XI^{e} siècle, ensuite le répertoire des oeuvres du XII^{e} siècle.

La publication de ce répertoire et l'enregistrement des textes sur bandes magnétiques, ainsi que la constitution de con-

cordances, de listes de fréquences et d'autres types de relevés, sont des opérations qui sont menées de pair.

Nous avons commencé par coder chacune des oeuvres depuis le VII^e^ jusqu au XII^e^ siècle inclus, en faisant la distinction entre les datations certaines et les datations incertaines, ainsi qu'à l'intérieur de chacune de ces catégories, la distinction entre oeuvres hagiographiques et oeuvres non-hagiographiques. Il est évident que certaines datations établies selon l'état actuel des recherches, vont être remises en question; nous le savons et avons déjà mis en place les programmes pour procéder aux transformations nécessaires.

Pour chaque oeuvre nous donnons les renseignements suivants: les « incipit », les éditions, les principaux travaux, les dates, des renseignements sur l'auteur, les sources ou des caractéristiques particulières, ainsi que le nombre de mots que compte l'oeuvre. Nous avons actuellement terminé les premiers traitements informatiques pour les oeuvres du VII^e^ au X^e^ siècle. Nous terminerons à la fin de cette année l'enregistrement des oeuvres du XI^e^ et XII^e^ siècle.

L'ensemble de ce très vaste corpus doit être lemmatisé; cette lemmatisation a déjà été effectuée pour un certain nombre d'oeuvres. Au fur et à mesure que nous avançons, nous réalisons un certain nombre de regroupements. Ainsi avons-nous déjà établi une concordance lemmatisée unique pour toutes les oeuvres hagiographiques belges du IX^e^ siècle, en y incluant, tout en les distinguant, les oeuvres du VIII^e^ - IX^e^ siècle et celles du IX^e^ - X^e^ siècle. On aboutit de la sorte à la constitution de divers types de « thesauri » en attendant, *in fine,* le « thesaurus » général des oeuvres narratives médiolatines de nos régions datées du VII^e^ au XII^e^ siècle. Ce vaste corpus devra être complété ultérieurement en y incluant des textes diplomatiques et en poursuivant le traitement au-delà du XII^e^ siècle.

P. Tombeur

Voici très brièvement certains traits caractéristiques des oeuvres médiévales que j'ai étudiées.

Comme vous le savez, tous les médiévaux citaient abondamment Aristote dans leurs oeuvres. La plupart du temps leurs citations trouvaient leur source dans les florilèges. On peut dire, par exemple, que saint Bonaventure a souvent cité Aristote dans son oeuvre à travers le florilège que j'ai étudié, ce qui montre bien l'importance de ces recueils pendant le moyen-âge. J'ai établi une concordance des citations contenues dans les *Auctoritates Aristotelis* et j'ai essayé d'identifier toutes les citations contenues dans le recueil. Les 3000 citations ont été identifiées, exception faite d'une cinquantaine. Pour rendre ces identifications facilement utilisables par le chercheur, j'ai réalisé à l'aide de l'ordinateur des tables d'identifications qui contiennent en regard de la citation du florilège son identification exacte avec la référence précise à une édition moderne.

Ces tables nous permettent de constater que beaucoup de citations attribuées à Aristote ne se trouvent pas dans ses oeuvres mais bien dans les oeuvres d'autres auteurs. Les médiévaux faisaient passer sous le nom d'Aristote nombre de phrases dont il n'était pas l'auteur. Dans ces cas-là, j'ai essayé de retrouver la source exacte de la citation.

Le second projet dont je m'occupe est le *Thesaurus Bonaventurianus*. Cette entreprise est vaste puisqu'il s'agit de traiter toute l'oeuvre de saint Bonaventure qui compte environ 3.000.000 de mots. Ce projet a été lancé il y a trois ans par Monsieur Wenin. Le premier volume de la collection a paru en 1972, le suivant paraîtra l'année prochaine.

Outre la concordance, l'index et les listes de fréquence, nous présentons dans ces volumes une étude des sources utilisées par saint Bonaventure. Pour analyser ces sources, j'ai essayé de distinguer les citations et les réminiscences, ce qui n'est pas toujours très simple dans les textes médiévaux. J'ai codé tous les mots qui appartenaient à une citation ou à une

réminiscence et ces codes apparaissent dans la concordance, ce qui permet de distinguer immédiatement le vocabulaire propre à Bonaventure et le vocabulaire emprunté à ses sources. Ce projet s'inscrit dans le cadre d'une recherche plus vaste concernant le latin scolastique. Nous essayons d'étudier la langue scolastique en tant que phénomène linguistique avec toutes ses caractéristiques propres: langue technique, langue de traduction... Dans le cadre de ce projet nous avons commencé un séminaire qui a pour thème l'étude du latin des auteurs universitaires du XIII^{e} siècle. Nous avons basé ce séminaire sur la comparaison de faits de langue chez deux auteurs systématiques proches: Thomas et Bonaventure. Nous avons pris des textes parallèles de ces deux auteurs et nous essayons de voir comment ils expriment des notions identiques. Le vocabulaire qu'ils emploient est souvent très différent.

J. Hamesse