Présentation des travaux du Centre de Recherche et d'Applications Linguistiques (C. R. A. L.) de Nancy

(Section des textes littéraires du Moyen Age français)

Le Centre de Recherches et d'Applications Linguistiques, en abrégé : C.R.A.L., est un Laboratoire de la Faculté des Lettres et des Sciences Humaines de Nancy (associé au C.N.R.S.). Il est issu de la fusion de deux groupes pré-existants : l'un établi à Nancy : le Groupe de Traduction Automatique de la Faculté, fondé en 1960 et dirigé par MM. Bourquin et Pottier (alors professeur à Strasbourg et maintenant professeur à la Sorbonne) et l'autre, le Groupe d'Analyse Automatique des Langues (G.R.A.A.L.) fondé à Strasbourg par M. G.-E. Weil et qui s'occupe de recherches bibliques et massorétiques. Le C.R.A.L. proprement dit a été constitué et reconnu par le C.N.R.S. le Ier janvier 1966. Il comporte quatre directions de recherches principales :

1) Traitement automatique des langues naturelles (traduction automatique) et de la documentation avec M. Bourquin (traduction automatique anglais-français), M. Pottier (traduction automatique espagnol-français, documentations linguistique et amérindienne) Mme Créhange (documentation scientifique et langage documentaire).

2) Département sémitique avec M. G.-E. Weil (analyse automatique de

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

textes hébreux, araméens, syriaques, ougaritiques, analyse paléographique de l'hébreu et documentation biblique automatique) et M. Atallah (analyse automatique de l'arabe).

3) Traitement automatique des documents historiques (histoire du Moyen Age) avec Mme Fossier.

4) Traitement automatique des textes littéraires français avec Mlle Naïs (textes médiévaux et du XVIe siècle), M. Mourot (stylistique littéraire) et Mlle Brayer (documentation littéraire de textes médiévaux).

Après cette courte présentation d'ensemble et à la suite de Mme Fossier qui a pris la parole ce matin pour sa section, je voudrais dans ma communication de ce soir vous exposer rapidement les travaux de mon équipe (textes d'ancien et de moyen français).

Notre expérience ne diffère pas sensiblement de celles qui ont déjà été décrites aujourd'hui même, aussi je pense que ma communication pourra sans inconvénient être assez brève. Comme la plupart d'entre vous, nous avons d'abord dû maîtriser assez péniblement des difficultés matérielles et nous présentons même la particularité d'être spécialisés dans la recherche sur ordinateur ... sans ordinateur. C'est-à-dire que, comme beaucoup d'utilisateurs littéraires nous n'avons accès à la machine que parcimonieusement et même occasionnellement.

Quelles sont les difficultés propres à nos recherches ? D'abord, l'établissement du texte. En effet, même lorsque les textes médiévaux ont reçu une (ou plusieurs) édition critique, celle-ci ne peut rendre compte exactement de l'état du texte dans les manuscrits. Or, lorsque l'on enregistre un

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

texte dans un ordinateur, il convient d'avoir le souci d'emmagasiner le maximum d'informations. Il nous a donc paru indispensable de partir des manuscrits et non des éditions modernes.

Je voudrais ici attirer l'attention des participants à ce colloque qui sont en majorité spécialistes de l'antiquité classique (ou des textes sacrés). Certes, leurs manuscrits présentent des variantes. Néanmoins leurs copistes visaient à la fidélité. Ce respect scrupuleux, religieux, du texte n'existe à aucun degré dans la transmission de la littérature médiévale en langue vulgaire. Les textes sont corrigés au gré de l'appartenance dialectale, de l'incompréhension (parce que la langue change), au gré aussi du goût des ateliers de copistes. Il n'est certes pas rare que deux manuscrits soient très proches, on peut dire à l'erreur matérielle près, mais si un texte est transmis par plusieurs manuscrits, même seulement un très petit nombre, les variantes sont souvent si nombreuses que l'on peut difficilement dire s'il s'agit de la même oeuvre ou de plusieurs versions de la même oeuvre. En particulier, pour un texte en prose, s'il y a, par exemple, un siècle de distance entre deux manuscrits, le second sera, pour le moins, la réécriture du premier texte en français du siècle suivant. Ce manque de certitude et de stabilité dans les textes a beaucoup conditionné notre travail et nous a inspiré nos principales directions de recherche.

Tout d'abord, contrairement à la plupart des participants à ce colloque, il faut bien dire que nous n'avons pas la lemmatisation pour principal souci. Bien sûr, nous rencontrons le problème, mais pour nous la fabrication d'index n'est qu'un sous-produit, et un instrument de travail dans nos propres recherches. Notre but était l'analyse linguistique du mot dans la phrase, l'analyse des énoncés grammaticaux. Nous partons donc de la phrase et non du mot. Le premier problème que je voudrais esquisser

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

devant vous est, en effet, celui-ci : fallait-il enregistrer le texte mot par mot ou en séquence ? Nous avons opté pour la deuxième solution, parce que lorsque l'on travaille sur un support magnétique, on peut toujours redécouper à volonté la phrase en mots, alors qu'il est beaucoup plus difficile de remonter du mot à la phrase. J'ajouterai à toutes les critiques désormais classiques de la notion de mot le fait qu'elle est encore un peu plus floue dans nos textes qu'ailleurs. Le jeu des agglutinations et des fausses coupes fait que le "mot" du manuscrit médiéval (l'écrit entre deux blancs) ne coïncide souvent pas avec le "mot" de sa transcription en édition moderne. On voit que nous nous heurtions à la nécessité d'enregistrer une double transcription du texte permettant de tenir compte de l'occasion de la double écriture du mot. Il fallait donc enregistrer le texte en séquence, sans se dissimuler la difficulté au niveau de la phrase également. Car la ponctuation des manuscrits médiévaux, si elle n'est pas arbitraire comme on le dit trop souvent, ne correspond absolument pas à nos usages modernes (surtout français). Nous devions donc redécouper en phrases significatives, mais nous avons également enregistré la ponctuation des manuscrits, ce qui permettra là encore d'étudier, si on le désire, la ponctuation mediévale.

La technique d'enregistrement, inspirée de celle de M. G.-E. Weil pour les textes bibliques, a été la suivante : des cartes perforées reproduisent le texte en séquence, les commentaires dont on a besoin d'assortir le texte sont ajoutées sur d'autres cartes qui correspondent aux premières, reprenant la technique des "zones" en cartes perforées. Actuellement, nous travaillons par blocs de trois cartes : une pour le texte dans un état d'édition semi-diplomatique (avec ponctuation moderne, résolution des abréviations, introduction des *j* et des *v* et des *é* en finale), une carte dite de contrôle paléographique qui comporte toutes les indications permettant

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

de reconstituer l'état réel du texte dans les manuscrits (code pour les parties abrégées, indication des agglutinations et fausses coupes, des fins de lignes, de la ponctuation du manuscrit, etc.), une carte d'analyse grammaticale. Cette dernière carte introduit l'analyse faite manuellement, mais le but est l'analyse automatique. L'analyse manuelle a permis de constituer les bibliothèques de bases, de désinences dont nous parlerons dans quelques instants.

Auparavant, il est nécessaire de dire de quel texte nous nous occupons. Il pourrait paraître souhaitable d'enregistrer le maximum de textes, mais l'exiguité de notre équipe est telle qu'il est impossible de l'envisager. Nous avons commencé par l'*Histoire de la Conqueste de Constantinople* de Villehardouin et il nous retiendra probablement encore longtemps. Les raisons qui ont guidé notre choix sont nombreuses et avant tout le désir de ne pas accumuler les difficultés occasionnelles, tout en présentant déjà, sous forme d'échantillon en quelque sorte, les principaux problèmes qui devaient se poser dans cette étude. Un texte en prose, non traduit, paraissait le meilleur point de départ pour une équipe destinée à ne pas pouvoir accroître très vite le corpus de son expérience. Un des manuscrits du texte, le manuscrit B de Faral, présente en outre une relative discrétion dialectale et aussi une certaine stabilité linguistique, ce qui était plus favorable au départ qu'un manuscrit qui aurait fait preuve d'une véritable anarchie (phénomène qui n'est pas rare). En outre, le texte de Villehardouin est l'un des premiers à pouvoir être daté avec précision (entre 1207 et 1213); ainsi était déjà supprimée l'une des incertitudes "classiques" des études d'ancien français : la date de composition; mais une autre demeure, car les manuscrits qui nous sont conservés sont tous postérieurs au moins d'un siècle à la composition et nous ne connaissons donc pas directement l'état original. Enfin, une recherche sur l'ancien

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

français implique que l'on puisse envisager une étude comparative de manuscrits d'un même texte et pour commencer s'il fallait plusieurs manuscrits, il n'en fallait pas trop : la tradition pour Villehardouin est constituée par six manuscrits, dont l'un au moins est complètement picard, plus deux éditions du XVIe siècle, dont l'une comporte un texte ancien et une juxtalinéaire en français du temps. On voit que les problèmes linguistiques étaient nombreux et divers sur une matière qui avait juste le format voulu pour une première expérience (le texte de Villehardouin atteint environ 45.000 mots).

Quelle a été notre démarche une fois un premier manuscrit du texte enregistré ? Nous avons d'abord entièrement codé les analyses grammaticales* pour pouvoir sortir rapidement des instruments de travail : index complet des formes, analysées, index inverse, index par classe de mots, ainsi qu'un index par occurrences. Mais le but de notre recherche est de pouvoir analyser les formes de manière entièrement automatique. La première étape de notre projet est l'analyse morphologique automatique. Elle est maintenant en bonne voie et nous pourrons en publier rapidement les premiers résultats. Je n'insiste pas, car les méthodes d'analyse morphologique commencent à être bien connues et nous n'avons pas vraiment innové en la matière.

La deuxième étape portera sur l'analyse syntaxique. Une première étude

---

* Notre code s'inspire à la fois du code de Liège et de celui de Besançon, exposé autrefois par M. Mitterand (Bull. d'Inform. du Laboratoire d'Analyse lexicologique, Besançon, n° 1 et n° 2, 1960). Voir aussi les art. de MM. Gougenheim et Bonnard, au n° 3, 1960, de la même publication).

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

porte sur le groupement nominal. Ajoutons que pour la recherche syntaxique nous pouvons profiter des travaux du groupe traduction automatique, autre section du C.R.A.L., qui s'occupe de ces problèmes depuis beaucoup plus longtemps que nous.

Mais je voudrais insister davantage sur ce qui fait à nos yeux l'originalité de notre tâche et montrer quelles options nous avons prises, là où le chercheur en informatique doit choisir. Le programme d'analyse morphologique, par exemple, comprend, nécessairement, une bibliothèque de "bases" et une bibliothèque de désinences. On peut constituer la bibliothèque de bases une fois pour toutes, à partir de dictionnaires, ou bien l'accroître au fur et à mesure, selon les textes étudiés. Nous avons préféré la seconde solution et voici pourquoi : une première raison, un peu accessoire, mais qui a son importance pratique, est que les variantes orthographiques dans nos textes atteignent de telles proportions qu'il est impossible d'espérer les trouver toutes enregistrées dans un dictionnaire quelconque. Or, pour un ordinateur, *commencier* et *conmencier*, par exemple, offrent deux bases différentes, de même *evesque* et *esvesque,* etc. Aussi, même en passant simplement d'un manuscrit à l'autre d'un même texte, on doit récrire en grande partie la bibliothèque des bases. La prévoir en entier dès le départ impose un travail à la main considérable, puis donne aux programmes de traitement des dimensions telles que l'on court à l'échec, même avec des ordinateurs de bonne taille de la 3e génération. Pour éviter de saturer trop vite les mémoires centrales, nous n'avons donc enregistré d'abord que les bases strictement nécessaires à notre premier manuscrit.

L'intérêt de cette méthode apparaît peut-être encore plus grand, si l'on envisage d'autres plans de recherche. En effet, notre but général est

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

d'arriver à définir la langue d'un texte, en elle-même d'abord, par comparaison avec d'autres textes ensuite. Or, nous n'avons de la langue de Villehardouin qu'une image déformée par les différents manuscrits. Sans vouloir entrer dans la discussion de savoir si nous pouvons atteindre une langue, ou s'il ne convient pas mieux de parler de la réalisation en discours, je me bornerai à deux remarques : 1) même le "discours" de Villehardouin nous est difficilement accessible, puisque nous ne pouvons que conjecturer son texte original; c'est de lui que nous nous occupons surtout et l'un des principaux buts de notre recherche est d'essayer de retrouver ce "discours", sous le travesti des manuscrits, ou au moins d'essayer d'apprécier si une telle entreprise est possible; 2) la tentation demeure d'essayer de passer du discours à la langue, ne serait-ce que sur le plan morphologique, c'est-à-dire qu'il faut trouver avec quoi compléter, par exemple, les nombreux trous dans le système morphologique, que laisse subsister le texte de la *Conqueste de Constantinople.*

On ne s'étonnera donc pas d'apprendre que nos recherches se dirigent beaucoup du côté statistique, en particulier de la recherche d'homogénéité. Dans quelle mesure deux manuscrits de Villehardouin sont-ils homogènes l'un à l'autre ? Dans quel autre texte peut-on légitimement aller chercher les formes verbales qui complèteront un système particulièrement lacunaire chez l'historien de la IVe Croisade (sa qualité d'historien n'est pas étrangère à ce fait : elle le fait recourir de préférence à la 3e et à la 6e personne et aux temps du passé et ne fournit presque rien pour le reste) ? Cette recherche de l'homogénéité est complexe et nous demandera beaucoup d'études mais nous pensons qu'elle pourra connaître de grands développements et permettre d'apprécier, par exemple, quels peuvent être les critères de critique interne pour l'attribution d'auteur et même s'il est possible de trouver des critères de critique interne pour résoudre ce point.

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.

Un autre domaine de la recherche statistique porte sur la théorie des sondages, qui a déjà fait l'objet d'un volume par M. et Mme Tomassone : *Recherches sur l'Application de la Théorie des Sondages aux Etudes de Linguistique* (Cahier du C.R.A.L., n° 3, 1969). Là encore les applications peuvent être nombreuses, notamment pour l'étude de l'ancien français. A ce niveau, nous ne disposons que d'un nombre fini d'échantillons, malheureusement pas choisis suivant les règles du hasard mathématique. La théorie des sondages devrait permettre d'apprécier quelles chances d'erreurs sont les nôtres pour la connaissance de l'ancien français. Quant aux langues modernes, elles devraient permettre de déterminer l'echantillonnage optimum permettant de rendre compte de ces langues. Il faut d'ailleurs reconnaître que l'analyse linguistique doit être menée concurremment pour préciser les buts de l'enquête. Ce n'est peut-être pas un des moindres avantages de cette méthode des sondages que de contraindre à préciser toujours les conditions et les buts des enquêtes.

On voit donc que, au-delà des problèmes matériels très nombreux que pose l'analyse automatique sur ordinateur, elle permet d'entrevoir d'immenses possibilités pour la recherche en linguistique quantitative.

Faculté des Lettres de Nancy
Centre de Recherches et
d'Applications Linguistiques (C.R.A.L.)

Hélène NAÏS

Extrait de la Revue (R.E.L.O.)
VI, 1 à 4, 1970. C.I.P.L. - Université de Liège - Tous droits réservés.