LE TRAITEMENT AUTOMATIQUE DE L'INFORMATION LEXICOLOGIQUE. REALISATIONS ET PROJETS

Cl. DUBOIS, Assistant à l'Université de Liège

Permettez-moi d'abord de remercier les organisateurs de ce congrès qui m'ont permis de présenter nos travaux.

Je me présente ici en tant que porte-parole de l'Institut de Lexicologie française de l'Université de Liège. Cet Institut, créé en 1963 et placé sous la présidence du Professeur Delbouille, s'est assigné pour tâche une étude approfondie du vocabulaire de l'ancien français. Le corpus sur lequel nous travaillons est limité aux textes français littéraires des XIIe et XIIIe siècles. Néanmoins, cela constitue une masse de documentation suffisante pour justifier le recours aux techniques électroniques. C'est en 1965 que nous avons mis à l'étude l'élaboration de concordances de nos textes. Nous avons choisi cette formule parce qu'elle nous paraissait réunir plus ou moins les avantages de l'index simple et ceux de la fiche lexicographique traditionnelle. Il nous semble actuellement que ce point de vue pourrait être sujet à de légères modifications : j'y reviendrai dans un instant.

Notre équipe compte en tout et pour tout cinq chercheurs. En outre, nous ne pouvons pas nous consacrer à plein temps aux activités de dépouillement parce que nos préoccupations premières sont des études linguistiques plus fondamentales telles que la sémantique, la syntaxe formelle, les systèmes suffixaux en ancien français et enfin la computational linguistics au sens propre.

Nos moyens matériels n'étant, en outre, pas très importants, nous avons dû essayer d'être les plus pratiques possibles.

La programmation de nos concordances est d'un type tout à fait traditionnel. Elle a d'abord été effectuée sur un IBM 7040. Actuellement et depuis un an nous utilisons conjointement un 360/44 et un 360/65. Ce matériel appartient en fait au Centre de Calcul de notre Université. Néanmoins, la programmation est assurée par nos soins.

Nous avons fixé la longueur du contexte de chaque forme à 130 caractères, soit la ligne machine. Cela garantit un contexte moyen de 3 octosyllabes, ce qui est assez lisible et assez peu souvent ambigu.

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

Nos concordances sont lemmatisées, c'est-à-dire que chaque forme du texte est ramenée à une forme canonique comme dans les dictionnaires. Cette lemmatisation suppose donc une analyse partielle des mots du texte. Nous employons pour cela une technique un peu particulière que je me permets de vous exposer avec plus de détails.

Nous avons envisagé avec attention la possibilité d'un système automatique de repérage des lemmes à partir de la morphologie. Et un certain travail préliminaire a été fait dans ce sens. Il nous est cependant vite apparu que le haut degré de variation tant graphique que morphologique de l'ancien français rendait l'approche de ce problème particulièrement ardue. Et, même si nous étions parvenus à écrire des programmes dans ce sens, il est vraisemblable que leurs rendements n'auraient pas été suffisants pour être raisonnablement amortis.

Nous utilisons donc pour le moment un système hybride qui consiste à faire proposer à la machine des hypothèses de lemmatisation basées non sur une reconnaissance mais sur une probabilité de reconnaissance. Nous avons introduit dans la machine un dictionnaire de base de 8000 entrées lemmatisées comprenant *d'une part* la forme standard de quelques milliers de mots parmi les plus fréquents choisis dans le Tobler-Lommatzsch et dans l'index de Chrétien de Troyes de Foerster.
D'autre part, nous y avons ajouté quelques milliers de formes fléchies ou déclinées de ces mêmes mots, mots dont la lemmatisation avait été effectuée manuellement lors de l'établissement artisanal de nos deux premières concordances. La comparaison des mots d'un nouveau texte à lemmatiser se fait par rapport au dictionnaire, forcément incomplet. Mais les formes du texte non représentées dans le dictionnaire seront néanmoins analysées et recevront, en guise d'hypothèse, la lemmatisation la plus probable, c'est-à-dire celle de la forme du dictionnaire la plus proche dans l'ordre alphabétique. Ce système n'est évidemment utilisable que dans la mesure où nous ne visons pas à donner une analyse de la forme mais seulement à attribuer un lemme. Bien entendu, il laisse inchangé le problème de l'homographie qui, de toute façon, ne pourrait trouver de solution qu'au niveau d'une analyse syntaxique, encore bien plus inaccessible que l'analyse morphologique.

Cependant, malgré son caractère quelque peu barbare, ce petit bricolage nous rend d'énormes services : il nous fournit actuellement 70 à 72 % d'hypothèses correctes. Ce pourcentage va sans doute s'améliorer encore dans la mesure où à chaque nouvelle concordance, nous introduisons dans le dictionnaire un certain nombre de formes nouvelles pour corriger les erreurs précédentes de l'ordinateur. Mais il va de soi que, jamais nous n'atteindrons ce paradis des lexicographes que serait le 100 % d'exactitude !

Tels qu'ils étaient, ces programmes du 7040 nous ont rendu de grands services. Cependant, lorsque s'est posé à nous le problème de la reconversion au nouveau matériel, nous avons choisi la voie la plus longue : c'est-à-dire non pas simplement accommoder nos programmes aux exigences du 360, ce qui au total eût été sans grands problèmes, mais profiter de l'occasion pour envisager une refonte fondamentale de notre système.

Le premier but vers lequel tendent nos efforts est l'amélioration du rendement. Les programmes que nous avons utilisés sur le 7040 étaient lents et, par voie d'immédiate conséquence, nous coûtaient cher ! Encore une fois il ne faut pas perdre de vue que nous sommes un Institut à vocation lexicologique essentiellement et que nous nous adressons à l'ordinateur pour qu'il nous aide et non pour qu'il constitue une charge. C'est pourquoi notre nouvelle programmation abandonne la mémoire bande magnétique et fait résolument appel aux mémoires à accès direct (mémoire centrale et disques). Du côté de la consultation du dictionnaire, par exemple, l'accès direct, si nous l'utilisons intelligemment, doit nous être extrêmement précieux et cela même sans craindre l'augmentation du nombre des en-têtes - crainte qui eut été pleinement justifiée avec les bandes magnétiques.

Le second point de concentration de nos efforts concerne l'obtention d'une plus grande souplesse, d'une plus grande polyvalence de notre système de programmation. Lorsque nous avons débuté, en 1965, nous avons décidé d'écrire des programmes qui donneraient des concordances, concordances dont nous avions en outre fixé les spécifications avec précision. Or, à la réflexion, nous ne sommes plus sûrs du tout que cette manière de voir les choses soit la meilleure car le lexicologue n'a pas besoin *que* de concordances, et même, on peut imaginer beaucoup d'espèces de concordances. Il y a surtout, pour nous lexicologues, une formule particulièrement intéressante et économique, c'est celle des concordances sélectives, conçues d'une manière suffisamment souple pour satisfaire les besoins des chercheurs, aussi différenciés soient-ils.

Or, en fait, pour obtenir toute cette diversité de résultats, les opérations de programmation se répètent bien souvent identiques à elles-mêmes pour le principe, et ne se distinguent les unes des autres que par de légères différences ou encore par l'ordre dans lequel elles sont effectuées. Aussi, sur la base de notre expérience antérieure, nous avons conçu un système formé d'une bibliothèque de petits sous-programmes, de modules, soumis à un certain nombre de paramètres et s'assemblant comme un jeu de construction, les programmes principaux se réduisant alors à une série d'appels de sous-programmes dans un ordre déterminé selon les besoins du moment. Le langage utilisé est le Fortran IV.

Nous espérons ainsi être capables de fournir au chercheur lexicologue, et cela dans un temps très bref, de la documentation "sur mesure" et nous débarrasser ainsi de la rigidité et de la monovalence (si on me permet le néologisme) des programmes précédents. Il est également prévu des "points d'accrochages" possibles pour un éventuel programme d'analyse morphologique si, dans l'avenir, une entreprise de ce genre nous semblait pouvoir être rentable. Car nous voulons absolument maintenir l'ordinateur dans ce rôle - selon un mot célèbre - de "libérateur de l'esprit", rôle dont il aurait parfois tendance à sortir !

Avant de terminer, je voudrais décrire très rapidement les caractéristiques principales d'un système automatique de documentation lexicologique *idéal,* ou du moins tel que notre expérience des six années précédentes nous le font imaginer et tel, certainement, que nous tenterions de le réaliser si les moyens matériels nous en étaient donnés.

En prenant pour principe de base la formule de la concordance traditionnelle, on étudierait la possibilité de mettre au point la programmation pour un ordinateur d'un *système de gestion* de la documentation lexicographique - une "banque de données" - qui s'inspirerait des systèmes employés en automatique documentaire (storage and retrieval). D'une manière générale, le système devrait avoir les caractéristiques suivantes.

1. *LE CORPUS.* Le systeme devrait pouvoir être interrogé soit sur un seul texte à la fois, soit sur un ensemble de textes préalablement désigné par le chercheur.

2. *POUR CHAQUE FORME-TEXTE,* outre les références aux oeuvres et un ensemble de renseignements à preciser prealablement (lemmes, analyses, traductions ou definitions, bibliographie, etc.), le systeme fournirait *un contexte* dont la longueur devrait être variable et mesurable en "unités" différentes : en nombre de caractères, en nombre de vers ou de lignes, en nombre de phrases ou portion de phrases (d'après des tests sur la ponctuation). Cette variation de la longueur du contexte devrait se faire selon deux types de critères différents.
   En premier lieu, selon des critères externes : la longueur est imposée par le chercheur, soit d'une manière uniforme, c'est-à-dire la même pour toutes les formes, soit d'une manière variable, le chercheur demandant alors des longueurs différentes pour chaque forme ou groupe de formes.
   En second lieu, selon des critères internes : la longueur du contexte varie alors suivant la catégorie grammaticale, par exemple (un plus petit contexte pour les mots dits "vides", un plus long pour les verbes, les substantifs, etc.), ou d'après la fréquence de la forme dans le corpus considéré (un contexte plus court pour les formes très fréquentes, plus long pour les formes rares).

3. *QUESTIONS - REPONSES.* Les renseignements précités devraient être fournis au chercheur dans les conditions suivantes.
   Exhaustivement, d'abord, pour un texte ou groupe de textes désignés. Nous retrouvons donc ici la formule de la concordance traditionnelle, telle que la propose notre Institut, par exemple. Mais d'autres types de questions sont imaginables. J'en citerai rapidement quelques-unes.
   On peut interroger le système sur la base d'une *liste de formes* donnée par le chercheur. Cette liste pourrait être *positive,* c'est-à-dire contenir les formes à propos desquelles on désire obtenir des renseignements. Elle pourrait également être *négative,* c'est-à-dire contenir les formes non-intéressantes et à ignorer purement et simplement.

   Au lieu d'une liste de formes, il doit être possible de donner une *liste de lemmes.* Le système fournirait alors les renseignements à propos des différentes formes appartenant à ces lemmes. Bien entendu, dans ce cas-ci également, la possibilité de listes négatives devrait être également prévue.

   On peut également concevoir l'interrogation du système sur la base d'une ou plusieurs catégories grammaticales. On obtiendrait une réponse avec classement sur les lemmes ou sur les formes. A côté des cri-

tères grammaticaux, on peut songer aux critères de fréquences. Par exemple, demander la liste de toutes formes et les renseignements y afférents, dont la fréquence, dans le texte ou groupe de textes considérés, est inférieure à 15.

Une autre base de sélection extrêmement intéressante est constituée par des critères contextuels. Par exemple : donner la liste des substantifs précédés ou suivis par l'adjectif X en tenant compte de l'intercalation possible d'un mot, deux mots, etc...
Enfin, il va de soi que le système fournirait un ensemble de tables et de relevés statistiques.

Pour l'élaboration d'un tel ensemble - notons d'ailleurs que la liste ci-dessus n'est pas limitative - les efforts, selon nous, devraient porter principalement sur les points suivants.

Les procédés de lemmage doivent, certes, retenir particulièrement l'attention. Il est nécessaire, pour des raisons d'économie de main-d'oeuvre, d'optimaliser la consultation du dictionnaire par une étude approfondie des systèmes de programmation permettant des manipulations de listes à haute vitesse (binary search, tableaux de lettres, etc.).

L'amélioration des temps de réponses aux questions exige également un examen attentif des techniques de programmation des mémoires a accès direct, des catalogues, des "directories", etc...

Enfin, vu la quantité des données inherentes aux problemes des lexicologues et des lexicographes, les problèmes de hardware que posent les operations d'input et d'output ne sont pas à négliger et une étude des avantages respectifs des differents procedes d'enregistrement et d'impression actuellement disponibles s'impose de toute évidence.

Je sais que l'énumération ci-dessus n'a rien d'original, qu'il est presque sûr que chacun des points a déjà fait, ici et la, l'objet de recherches et de réalisations. Mais ce qui me paraît extrêmement important, c'est que les centres de recherches lexicologiques puissent disposer *de l'ensemble* des possibilités que je viens d'évoquer rapidement.

Et a ce propos, d'ailleurs, je voudrais joindre ma voix à celles, déjà nombreuses - pensons notamment à l'Organisation internationale pour le Traitement des Langues anciennes par Ordinateur créée par le Professeur L. Delatte - qui reclament instamment que des efforts de coordination soient faits pour éviter la dispersion des forces et la multiplication inutile des programmes.