QUELQUES PROBLÈMES ET SOLUTIONS MÉTHODOLOGIQUES

par BERNARD QUEMADA

Mes chers collègues, vous me permettrez de profiter de l'avantage de l'âge pour limiter mon court exposé à quelques aspects particuliers des travaux que je poursuis maintenant depuis quelques 18 ans. J'ai en effet eu l'occasion de les décrire de manière générale dans de nombreux articles et conférences. Je vais donc me borner ici à quelques vues synthétiques pouvant permettre de dégager quelques thèmes de réflexion et peut-être de rendre service. J'ai constaté hier avec beaucoup d'intérêt la passion qui animait certains moments du débat. Je crois, qu'en fait, cette passion s'exprime à propos d'un certain nombre de points fondamentaux sur lesquels nous devons essayer de nous entendre. Pour schématiser le tout, disons ceci: doit-on travailler *exclusivement* ou *partiellement* avec des machines? Les travaux sur machines, qui naturellement supposent l'exhaustivité, sont-ils préférables, plus sûrs, que les travaux qui mettent en cause l'objectivité des « travailleurs manuels » (très intellectualisés d'ailleurs)... Finalement, qui peut le plus et le mieux?

Je crois qu'il faut poser ici un point essentiel et je crois que l'on était pratiquement sur le point d'admettre, en se séparant hier soir, que chaque formule, chaque démarche peut beaucoup à condition de savoir ce que l'on veut faire et pourquoi on le fait. On ne sortira jamais d'une problématique hasardeuse, si l'on ne veut pas poser ceci comme point fondamental.

Je vais prendre quelques exemples de travaux que j'ai

réalisés, ou qui sont en cours d'élaboration, pour justifier les approches que nous avons adoptées. Par exemple, nous avons entrepris deux grands chantiers qui restent ouverts, multidimensionnels, l'un que nous appelons '*Fichier historique du vocabulaire français*', l'autre que nous appelons '*Inventaire du néo-français*'. Dans le *Fichier historique* il s'agit de rassembler, avec tout ce qu'il y a de risques d'arbitraire et de hasard, les premières attestations des « mots » français, pour dater leurs sens et le développement de leurs emplois, (et dans « mot » nous incluons les expressions, locutions, etc.). Vous imaginez combien tout ceci ouvre de perspectives immenses... Nous travaillons à cela avec une équipe qui est répartie à travers le monde; en fait nous recevons surtout les trouvailles de chacun (car le travail n'est jamais tout à fait systématisable), et nous les intégrons dans un fichier général qui est, bien sûr, très disparate. Parallèlement à cette première source de données, nous avons relevé, depuis douze ans que nous avons commencé ce travail, tous les témoignages qui figurent dans les dictionnaires historiques et étymologiques, et nous les avons perforés avec tout ce qu'il peut y avoir de contradictoire d'un dictionnaire à l'autre, et en suivant avec beaucoup d'attention les rééditions pour inclure les nouveaux éléments qui sont donnés par ces dictionnaires. Chaque année nous pouvons ainsi inclure plusieurs milliers de nouvelles données dans le répertoire de base.

Ce travail avait débuté suivant les méthodes traditionnelles manuelles, sans penser — et pour cause — aux possibilités offertes par les machines. Puis nous sommes passés aux cartes perforées, et aujourd'hui, à l'ordinateur.

Mais ceci est dans l'ordre des choses, car il fallait commencer par la phase manuelle pour la première collecte avant de recourir à la machine. Or maintenant, pour bien gérer l'entreprise, nous avons recours à elle. Mais j'ai dit: pour gérer l'entreprise; non pour constituer le stock des données fondamentales. Il y a une stratégie à imaginer pour certaines opérations

où l'on doit commercer à la main et où l'on continue seulement à la machine. L'ordinateur n'effectue pas ici les grands travaux, mais il intervient à sa place pour faire naître beaucoup de questions. Sur le travail fondamental entrepris au début à partir des dictionnaires historiques, par exemple, nous avons trié les observations et les éléments documentaires donnés par ces dictionnaires et nous les rangeons, par exemple, par texte d'origine. Nous découvrons ainsi que beaucoup de ces dates sont données d'après des éditions des XVIII[e]-XIX[e] siècles complètement déjugées aujourd'hui. Pour nous permettre de faire apparaître de nouvelles séries de datations, il suffit d'avoir recours à une édition récente plus valable.

Il est clair aussi que la recherche de nouvelles datations (qui sera faite toujours de manière assez aléatoire) gagne beaucoup à partir maintenant de listages qui donnent le dernier état des connaissances. Le listing mécanographique permet très rapidement de faire le point de l'état des connaissances.

Pour le deuxième cas, le problème est encore plus énorme. L'une des missions confiées à notre CENTRE D'ETUDE DU FRANÇAIS MODERNE ET CONTEMPORAIN est en effet d'observer le français actuel dans toutes ses mutations, évolutions, transformations, etc. Faute de personnel permanent suffisant, nous ne pouvons pas avoir un programme systématique. Nous faisons donc encore une fois appel à tous et nous avons animé, dans des universités des pays francophones et des pays non-francophones, des équipes d'observateurs avec une participation importante fournie par des étudiants. Par exemple à l'Université de Saarbrück et de Montréal, des équipes travaillent depuis plus de huit ans à cette opération et l'on y élabore, toujours à la main, des fiches donnant les premières informations sur des « mots » qui ne sont pas repérés dans les dictionnaires. Nous avons actuellement constitué un stock de 500.000 documents (manuels) normalisés. A côté de documents textuels proprement dits, nous avons là des photographies d'affiches qui ont été

prises dans la rue; car nous ne pouvons ignorer des informations sur la langue des affiches, de même que des reproductions de dessins avec une légende, parce que la légende contenait la création lexicale qui nous intéresse.

Pour le moment, nous n'avons pas le moyen de mettre tout cela sur ordinateur; mais tout est prévu pour que, dans un deuxième temps, cela se fasse. Nous sommes actuellement en train d'élaborer des bordereaux pour le matériel normalisé: nous l'avons traité à la main, classé à la main etc. mais nous allons faire dès 1975 des fiches récapitulatives qui passeront sur ordinateur pour la gestion de l'ensemble. Gestion qui nous permettra, par exemple, de suivre les développements dans toutes les directions de ces divers néologismes (nous avons parfois 150, 200 témoignages sur le même phénomène, des témoignages qui varient dans le cadre des 10 dernières années que nous observons), mais qui nous permettra surtout de *poser des questions.* Le plus souvent nous ne savons rien de ce mot, or il est clair qu'il existe en français depuis 10 ans: nous pouvons demander des informations à ce sujet. Encore une fois, c'est alors avec la machine que nous travaillerons. Ceci pour montrer simplement l'aspect complémentaire des démarches.

Par ailleurs, il est évident que ce fichier ne peut pas rester tel quel. Cette entreprise problématique ne peut pas être sans déboucher sur une initiative plus systématique, initiative qui se matérialisera vraisemblablement par un *Dictionnaire du Français de la fin du XX^e siècle.* Nous sommes très préoccupés à cet égard, mais nous n'avons pas encore arrêté de solution. Ce dictionnaire hypothétique, si nous le réalisons un jour, nous le ferons vers l'an 1985 ou 90. Pour cela nous joindrons, à l'*information aléatoire* dont nous disposons déjà, une *information systématique* à réunir ultérieurement. Il faudra donc, et c'est un peu le problème que vous abordiez hier, M. Gregory, imaginer une technique qui permette de reprendre, dans une perspective permettant une exploitation cohérente et commune,

des documents exhaustifs systématiques aussi bien que des dépouillements d'échantillons de textes, et que le matériel disparate rassemblé en ce moment. L'avenir, tel que nous l'imaginons aujourd'hui, nous conduit à envisager de constituer une *Banque de Données Lexicales sur le Français Contemporain.* Toutes les justifications ultérieures dépendront essentiellement de la problématique que nous allons adopter dès maintenant. D'où l'importance de réflexions du type de celles qui sont conduites ici aujourd'hui. C'est le point où nous sommes arrivés dans cette perspective. C'est ce que je voulais vous dire à propos de la première tranche de nos travaux, des travaux *sélectifs* mais que je qualifierai d'*ouverts*.

A côté de cela, il faut évoquer deux autres travaux que j'appellerai *systématiques*, et par là-même, *clos*.

Le premier exemple sera l'ensemble de recherches que nous poursuivons sur des répertoires lexicographiques. Pour satisfaire un rêve ancien, j'ai ouvert il y a maintenant plus de 15 ans un chantier immense visant à constituer un *Trésor des Dictionnaires anciens* en tirant profit des possibilités des machines mécanographiques — on ignorait encore l'usage des calculatrices. J'ai donc commencé par mettre sur carte, en utilisant les méthodes de l'analyse graphique interprétable par photolecture, le contenu des trois grands dictionnaires du XVII[e] siècle: RICHELET, FURETIÈRE et l'ACADÉMIE. Ces éléments constituaient la base d'une somme de données (déjà une « BANQUE »!) que nous pouvions tenter de consulter grâce à la machine, a partir des catégories d'informations qui se trouvaient recensées. De cette première expérience, que nous avons laissée en suspens car elle était beaucoup trop onéreuse pour les moyens dont nous disposions alors, nous avons tiré deux conclusions. La première est que l'entreprise, grâce à la mécanisation sans cesse plus performante, est parfaitement réalisable compte tenu de moyens importants — mais les résultats seront largement à leur mesure; la seconde, que le travail préparatoire (choix des répertoires, analyse des

données textuelles, définitions des normes et des types, etc...) était le gage essentiel de la valeur des résultats eux-mêmes.

A partir de ces « certitudes », les démarches préparatoires ont pu être développées sur plusieurs plans avec une confiance considérablement accrue. J'ai entrepris une quête minutieuse de tous les dictionnaires, glossaires, vocabulaires, lexiques, etc... du XVI[e] à nos jours pouvant offrir des informations lexicographiques en français. Cela a déjà donné une grosse bibliographie de plus de 20.000 ouvrages pour la seule période 1500-1863 qui doit paraître l'an prochain. Mes collaborateurs du CNRS préparent la matière pour la période suivante. J'ai parallèlement mis en route un programme de réédition sur micro-fiches des dictionnaires les plus importants dans le cadre d'une réalisation plus générale, les *Archives de la linguistique française,* qui doit compter à son terme quelques 1.500 titres et dont 400 sont déjà actuellement diffusés.

C'est à l'exploitation de ce capital disponible, grâce à l'informatique, que nous pensons très activement maintenant. En attendant les machines à lire directement les textes anciens imprimés, que va-t-on perforer? La constitution d'une *nomenclature des nomenclatures* nous a paru, après des discussions avec nos amis du TRÉSOR DE LA LANGUE FRANÇAISE, trop peu indicative car pour tous les dictionnaires (les anciens en particulier) de très nombreux mots cachés viennent fausser les inventaires directs des formes vedettes. Nous avons mis sur fiches, et sur bandes magnétiques la nomenclature du GRAND LAROUSSE ENCYCLOPÉDIQUE en 11 volumes. Nous avons d'abord trouvé plus de 120.000 vedettes. A partir de ces 120.000 vedettes, en analysant le contenu des articles, nous avons créé à peu près 500.000 items. Mais nous avons surtout repéré (sans être formels sur l'exhaustivité) 85.000 mots cachés, c'est-à-dire qui ne sont pas dans la nomenclature. Nous démultiplions donc les travaux en ce domaine et nous faisons appel à des étudiants pour cela, faute d'avoir des crédits pour un travail systématique. Les résultats

seraient encore plus frappants si nous avions la possibilité de traiter sur ordinateur la totalité du texte des dictionnaires. Un de mes élèves, excellent connaisseur de NICOT, est en train de préparer un tel programme pour analyser systématiquement le *Thresor de la langue françoise* de 1606 grâce à l'ordinateur de l'Université de Toronto qui lui fournira les concordances les plus diverses de toute la matière lexicographique. Par ailleurs, nous avons terminé la préparation d'un volumineux *Index français-latin* du vocabulaire (et non seulement de la nomenclature) des lexiques latin-français du Moyen-Age publiés par Mario Roques. Ce travail vise à rendre accessibles, par une présentation dans l'ordre alphabétique, tous les mots français qui étaient contenus dans le corps des gloses (synonymes, énoncés définitoires, développements encyclopédiques) et que la disposition originale rendait pratiquement « insaisissables ». Ce travail n'a pu être réalisé qu'à la machine — compte tenu, bien sûr, d'une importante préparation manuelle —. Et il serait très facile d'en multiplier d'identiques sur le modèle de nos expériences de « retournement » des dictionnaires bilingues que nous avons présentées il y déjà de nombreuses années.

Le dernier groupe de mes observations portera sur l'élaboration lexicologique et lexicographique des matériaux qu'il est devenu aujourd'hui très commun d'établir avec des machines. Il s'agit là aussi d'une de mes préoccupations essentielles.

J'assume à ce jour la responsabilité d'avoir fait perforer sur cartes quelques 18 millions de mots, dont l'essentiel a été réalisé pendant une courte période (moins de 5 ans). Sur cet ensemble, la préparation des documents de base (index des mots et inventaires statistiques) a été immédiatement effectuée, en apparence sans problèmes majeurs. Mais quand on y regarde de plus près, avec le recul du temps, on est vite tenté de se poser des questions qui ne nous venaient pas à l'esprit encore hier. Peut-on tomber directement sur des normes de traitement qui assurent aux résultats obtenus une portée qui dépasse la seule exploitation prévue

à l'origine? Celles qui retiennent ici et là les chercheurs seront-elles harmonisables? Il semble facile de soumettre le plus grand nombre de textes à un programme d'indexation automatique, mais qu'en tirerons-nous en réalité si les unités de traitement (mots et lexies) ne sont pas définies, si les postulats de la lemmatisation ne sont pas arrêtés en toute clarté, etc...? Le passage de la documentation proprement dite (élaboration des index et des concordances) à l'établissement des énoncés lexicographiques est généralement peu ou pas abordé par les spécialistes actuels. C'est pourtant là un problème capital auquel nos échanges de vues devraient accorder une place importante. J'oppose pour ma part, afin de préciser deux démarches distinctes bien que devenues complémentaires, les approches *lexicographiques*, d'une part, et *lexigraphiques* de l'autre. Ce second terme renvoie à tout ce qui concerne l'établissement de répertoires d'unités lexicales en contexte (essentiellement les multiples sortes de concordances que nous savons élaborer aujourd'hui avec la machine) pour lesquels les méthodes de la linguistique fonctionnelle peuvent être d'un très grand secours. A ce titre, les travaux qui sont entrepris pour la réalisation d'un prototype de *dictionnaire d'auteur* (il s'agit du dictionnaire des oeuvres de MOLIÈRE) devront nous offrir la matière pour illustrer cette conception. Le respect absolu de l'exhaustivité, associé à l'affichage minutieux des caractéristiques contextuelles d'emploi, sont à la base des principes méthodologiques que nous voulons suivre. Sur le même modèle, des *dictionnaires d'oeuvres* compléteraient les formules principales des réalisations *lexigraphiques*.

A l'inverse, les réalisations *lexicographiques* se caractériseront par une démarche sélective, économique parce que dominée d'abord par des impératifs pratiques (ce qui ne la condamne pas absolument à nos yeux, mais lui confère un caractère para-linguistique évident) et menée au prix d'un sérieux indispensable dans les choix, d'une manière plus « efficace » en fonction d'hypothèses de travail clairement établies. A ce titre, ce sont les pra-

ticiens et les usagers qui interviendront en premier lieu pour en définir les caractéristiques. Mais on est en droit alors de s'interroger sur le bien fondé des dépouillements systématiques sous lesquels, faute de critères automatiques de sélection, l'auteur de définitions risque d'être enfoui.

J'espère que ces considérations ne vous paraissent pas trop déroutantes. Elles visent à préciser au seuil de ces débats les principaux problèmes que l'usager des machines ne doit pas perdre de vue: complémentarité des phases manuelles et automatiques à définir très précisément, certains travaux ne pouvant aujourd'hui encore être entrepris qu'à la main; nécessité de bien délimiter l'usage des matériaux que les facilités offertes par les machines risquent de nous faire accumuler pour un usage problématique.