Meta
Journal des traducteurs
Translators' Journal

META

Bilan des applications de l'informatique aux études lexicologiques

Bernard Quemada

Volume 18, Number 1-2, mars 1973

Actes du deuxième colloque international de linguistique et de traduction. Montréal, 4-7 octobre 1972

URI: https://id.erudit.org/iderudit/003883ar
DOI: https://doi.org/10.7202/003883ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Quemada, B. (1973). Bilan des applications de l'informatique aux études lexicologiques. *Meta*, *18*(1-2), 87–102. https://doi.org/10.7202/003883ar

Bilan des applications de l'informatique aux études lexicologiques

Notre président vient de vous dire les préoccupations que nous partageons avec M. Zampolli de présenter devant vous des communications devant traiter deux sujets à peu près identiques. Il se trouve que je suis le premier à parler ce matin, je me permettrai donc de simplifier et de schématiser mon exposé le plus possible. J'espère ainsi qu'il pourra servir d'initiation à ceux qui ne sont encore pas très familiers des problèmes d'informatique, et, pour ce qui est des aspects plus techniques pouvant intéresser les spécialistes, d'introduction aux compléments que M. Zampolli aura le loisir de développer ensuite.

Le titre que j'ai donné à ce petit exposé : « Bilan des applications de l'informatique aux études lexicologiques » me paraît pouvoir se passer de plus d'explications. Disons seulement que j'ai pensé qu'il ne serait peut-être pas inutile qu'à l'occasion d'un colloque sur la lexicologie contemporaine l'on tentât de dresser un état des acquisitions les plus récentes dans ce domaine, à la lumière de quinze années d'expériences diverses menées un peu partout dans le monde, alors même que les théories linguistiques et lexicologiques, d'un côté, les équipements et les programmes, de l'autre, évoluaient très rapidement. Il est bien évident qu'un tel inventaire critique, même s'il reste très général, est étroitement dépendant de l'histoire — fort brève d'ailleurs — des applications et des matériels eux-mêmes. Je me suis trouvé être vers 1955 le premier linguiste français à s'intéresser aux possibilités qu'offraient les techniques de traitement des données linguistiques sur machines, ce que j'ai fait entre-apercevoir aux participants du Colloque de Strasbourg de 1957 [1]. Cette date marquera le début de cette rétrospective, en quelque sorte, celle de la préhistoire de l'utilisation des ordinateurs en lexicologie. Je prendrai comme limite terminale une date très proche de nous : il y a quelques jours, j'ai eu l'occasion de réaliser pour un séminaire international tenu à Pise au Centro nazionale de calcolo electronico une étude sur les techniques de pointe immédiatement exploitables. Il s'agissait de répondre à une demande de l'Accademia della Crusca, formulée au nom de tous les responsables des grands dictionnaires historiques actuellement en cours de réalisation, et pour lesquels le recours à la mécanisation ne s'est pas traduit que par des avantages. Les immenses possibilités documentaires des ordinateurs ont permis la constitution de dépôts de

1. Voir les *Actes du Colloque sur les orientations actuelles de la lexicologie et de la lexicographie romanes (Strasbourg, 1957)*, Paris, C.N.R.S., 1961. On se reportera sur nos premières expériences aux divers rapports publiés dans les *Cahiers de lexicologie*, I (1959) à IV (1962).

données d'une richesse inimaginable il y a seulement vingt ans. Mais l'ampleur de cette information pose aujourd'hui des difficultés d'un autre ordre, si importantes que l'on s'interroge parfois sur les progrès réels qui ont été ainsi acquis. Une nouvelle réflexion nous a paru s'imposer à ce terme. C'est donc bien un bilan que je vous propose, mais accompagné aussi de suggestions dont je ne voudrais pas dissimuler les ambitions réalistes.

C'est pour le principe que je ferai remarquer ici entre parenthèses que, dans tout cet exposé, *lexicologie* et *lexicographie* seront encore très étroitement associées. Comme dans le passé des réalisations linguistiques, la *lexicographie* science-technique des dictionnaires va devancer en apparence la *lexicologie stricto sensu* dans le recours aux possibilités de l'automatisation. Mais je crois que nous sommes tous d'accord pour reconnaître qu'il ne peut y avoir de barrière étanche entre les deux démarches de l'inventaire et de la description des « mots ». C'est sans doute parce que la *lexicographie,* domaine d'application, et par là susceptible de bénéficier de moyens matériels plus importants, a pu disposer de sommes souvent importantes lui permettant l'utilisation des ordinateurs. Il va sans dire que toutes les branches de la *lexicologie* en bénéficient à leur tour, et ceci dans des conditions d'autant plus satisfaisantes que les réalisations lexicographiques s'inspirent elles-mêmes de plus en plus des réflexions méthodologiques (linguistiques) les plus récentes.

La *lexicologie* est le domaine de la linguistique où l'utilisation des machines a pris assez vite une place importante et qui ne cesse encore de s'accroître. La traduction automatique a peut-être une histoire un peu plus longue, avec ses espoirs déçus et un retour en arrière, très significatifs des expériences du premier âge. En revanche, dans la filière des services que la mécanisation était tout particulièrement apte à rendre aux disciplines lexicologiques, la progression a été constante, tant sur le plan qualitatif que quantitatif. Elle a atteint aujourd'hui des résultats spectaculaires, et à bien des égards réconfortants pour tous ceux qui en assurent les débuts discutés.

Si l'on se reporte aux années 1950-1955, la lexicologie européenne semble se développer alors à partir de trois lignes de force, expression de tendances complémentaires dont l'influence se manifestera très rapidement :

1. *La lexicologie socioculturelle,* la plus ancienne, étroitement liée à la tradition linguistique essentiellement historique. Elle se présente, sous la plume de certains de ses protagonistes, comme une discipline « sociologique », lorsqu'elle poursuit l'analyse des « mots témoins de l'histoire ». Pour affirmer la rigueur de ses démarches, elle accentue les exigences de la documentation : seule une information « totale » — pour ne pas dire « exhaustive » — permettra d'identifier *mots clés* et *mots témoins* suivant les hypothèses de G. Matore. *L'inventaire général de la langue française,* lancé en 1936 par Mario Roques pour regrouper des emplois « caractéristiques » des mots français n'est plus à la hauteur de la tâche, puisque ce sont des programmes d'exploration systématique de toute la presse française de 1830 à 1848, par exemple, que l'on nous propose. Tous les travaux effectués à partir de mots choisis plus ou moins intuitivement doivent être tenus pour suspects. Le problème des moyens nécessaires pour atteindre cette

nouvelle dimension de l'information est posé : c'est d'abord dans cette perspective que j'ai songé à utiliser les machines.

2. *La lexicométrie et la statistique lexicale.* C'est vers ces mêmes dates que, pour le domaine français, les travaux de P. Guiraud montrent l'importance que l'on est en droit d'accorder à l'interprétation des données quantitatives, et tout spécialement à celles fournies par les statistiques de vocabulaire. Restait à leur apporter les possibilités de traitement à l'échelle où ces problèmes demandent à être envisagés et avec la précision qu'ils supposent. L'établissement des index à la main, les comparaisons entre des listes trop rares entraînaient erreurs ou rapprochements incongrus. Tout un arsenal documentaire était à créer au plus vite si l'on voulait poursuivre un développement sérieux. Il l'a été grâce aux machines et la lexicométrie a enfin pu prendre son essor véritable.

3. *La lexicologie fonctionnelle,* inspirée de la linguistique fonctionnelle structurale, va faire aussi, à peu près simultanément, son entrée et prendre la tête des méthodes de description lexicologiques. Si pour elle, les mots (et surtout leurs sens) ne peuvent être appréhendés qu'à partir des emplois, les recherches ne peuvent alors progresser qu'à partir de compilations complètes des collocations ou distributions. Aux *index* des mots (suffisants pour les statisticiens d'alors), vont s'associer dans cette nouvelle perspective les *concordances* contextuelles : à la main elles étaient irréalisables (celles de la Bible, au XVIII[e] siècle, ont demandé des années à des communautés). Ici encore, le secours de la mécanisation a été déterminant.

Il convient aussi de mettre en parallèle à ces options majeures les grandes réalisations (ou projets) lexicographiques poursuivies ou mises en place pendant la même période. Leurs impératifs particuliers ont été très favorables au développement des applications de l'informatique. On sait que les dictionnaires reflètent, avec un décalage variable, les options méthodologiques de la lexicologie et, par conséquent, de la linguistique contemporaines. Ceux qui vont surtout nous occuper sont les grands dictionnaires historiques nationaux (Florence, Leyden, Madrid, Gottingen, Édimbourg ou Nancy) : bien qu'ils continuent dans la seconde moitié du XX[e] siècle des projets européens qui remontent souvent au siècle précédent, ils n'en portent pas moins la marque de perspectives linguistiques récentes. Or, à tort ou à raison, mais fidèles en cela aux intentions primitives, leurs responsables se montrent surtout soucieux d'accumuler la documentation la plus riche. Comme ces grandes entreprises disposent, relativement, de beaucoup plus de moyens que la recherche universitaire, ce sont elles qui ont mis sur pied les premières, sur une très vaste échelle, l'élaboration de masses documentaires (ou complément d'un font ancien) en faisant usage de l'informatique.

Ainsi, considérations théoriques d'une part, besoins pratiques de l'autre, débouchaient dans un premier temps sur des *problèmes de documentation* : on ne sera donc pas surpris que ce soit dans cette perspective que les moyens offerts par l'informatique aient été tout d'abord mis à contribution.

Avant de parler des réalisations, il faut dire quelques mots des équipements, dans la mesure où plusieurs de leurs possibilités ont directement conditionné (inspiré ?) les applications qui en ont été faites.

Les machines auxquelles les lexicologues pouvaient faire appel vers 1955 ont beaucoup évolué au cours de ces quinze dernières années. Il n'est pas sûr que, dans cette course aux plus grandes performances, toutes les possibilités des divers équipements proposés aient été sérieusement reconnues, compte tenu de la relation coût-rendement. En règle générale, chaque performance supplémentaire allait éveiller un appétit correspondant chez les chercheurs imaginatifs et ouvrir de nouvelles perspectives. Pour prendre le plus « gros » exemple, disons que si l'ordinateur n'était pas né, le *Trésor de la langue française,* qui a pu grâce à lui rattraper et dépasser un retard de plus de cinquante ans de travaux préparatoires (sa documentation « électronique » de 70 millions d'occurrences textuelles a été réunie en 6 années environ), n'aurait certainement pas été entrepris.

Pour simplifier au maximum, je rappellerai seulement qu'il convient de distinguer deux catégories fondamentales de matériel : les *machines mécanographiques,* d'une part, et les *ordinateurs,* de l'autre. Les premières traitent des informations à partir de cartes perforées qui vont servir pendant toutes les opérations sur les diverses machines comme support des données. Les seconds traitent des informations « électroniques », qui peuvent être introduites à partir de cartes perforées, mais qui sont au préalable transformées en points magnétisés sur bandes et disques magnétiques, avec toutes les servitudes que cela suppose au niveau des modes et des temps de traitement ; les seconds offrent des capacités de stockage des données immenses (mémoires), de multiples possibilités d'automatisation de traitements particuliers (programmes) et surtout des vitesses sans cesse accrues (certaines opérations vont de 10 000 à 100 000 fois plus vite sur un ordinateur que sur une machine mécanographique).

À partir de ces indications générales, la tentation est grande aujourd'hui de dénigrer les équipements mécanographiques. Il convient pourtant d'observer avant d'aller plus loin que certains perfectionnements qui leur ont été apportés au cours des années en ont fait un matériel qui, s'il ne peut évidemment pas rivaliser avec les ordinateurs peut les compléter très avantageusement, compte tenu de leurs spécifications respectives. C'est donc à la fois pour des raisons techniques, pratiques et économiques que je crois à leur avenir s'ils sont judicieusement employés. Je vais donc sommairement vous les décrire en premier lieu. Des auditeurs ayant beaucoup insisté sur leur ignorance en cette matière, je vais vous proposer deux schémas très élémentaires :

1. *Les moyens mécanographiques* peuvent être présentés en trois sous-ensembles :

a) *Les cartes mécanographiques,* établies à l'entrée du système, sont les supports des informations alphabétiques ou numériques que l'on va traiter, accompagnées de données qui les explicitent. Elles sont le plus souvent perforées à l'aide d'une machine à clavier dactylographique.

b) *Le circuit des machines* sur lequel les cartes vont être successivement traitées comprend principalement : la *traductrice* qui imprime en clair sur les cartes les signes correspondant aux perforations, les *trieuse* et *interclasseuse* qui classent par ordre alphabétique ou numérique (avec possibilités de comptages statistiques),

la *reporteuse* qui imprime sur une carte des informations contenues sur une autre, la *reproductrice* qui reproduit sur une carte les perforations (toutes ou partie) contenues sur une ou plusieurs autres cartes, la *tabulatrice* qui imprime sous forme de listes ou de tableaux les fichiers mécanographiques, en faisant divers types de calculs ou suivant diverses règles d'édition, etc.

c) *Les archives mécanographiques* sont constituées par le stockage cumulatif des fichiers mécanographiques déjà élaborés. Elles sont susceptibles d'être consultées et surtout progressivement complétées par de nouvelles données, souvent tirées de l'analyse d'informations enregistrées au cours d'étapes antérieures.

Un tel système révèle aujourd'hui, à tous ceux qui connaissent des techniques plus récentes, des insuffisances évidentes : étant lié à des procédés mécaniques, il est d'abord beaucoup plus lent que les procédés électroniques. Le nombre de caractères distincts qui peuvent être traités sur ces diverses machines est insuffisant pour respecter la totalité des signes graphiques d'une langue comme le français. Des aménagements sont alors nécessaires et aboutissent parfois à sacrifier des éléments indispensables. Les principes même de fonctionnement des machines, entraînent l'impossibilité pratique d'explorer les environnements des unités lexicales, etc. Mais ces machines offrent en revanche des ressources appréciables qui ont déjà fait largement leurs preuves. Elles sont lentes, c'est vrai, mais leur vitesse est suffisante dans bien des travaux d'étendue moyenne, compte tenu des possibilités courantes des chercheurs pour l'interprétation des résultats. Elles sont relativement peu coûteuses, très faciles à utiliser, et, de ce fait, peuvent être confiées à des usagers non spécialistes (étudiants et chercheurs). Leur implantation auprès des utilisateurs permet à ceux-ci de suivre chaque étape du travail, d'intervenir très facilement et d'avoir toujours une vue concrète, tangible de l'élaboration des documents, ce qui confère aux machines mécanographiques des avantages pratiques autant que psychologiques non négligeables. Et il convient de signaler, en outre, que depuis 14 ans ces systèmes ont bénéficié de nombreux perfectionnements (nouveaux dispositifs d'entrée des données : marques graphiques portées à la main qui commandent les perforations automatiques, reports xerographiques ou insertion d'un microfilm — une ou plusieurs pages de contexte — sur une carte ; accélération ou aménagements nouveaux de certaines fonctions, comme le tri par lecture optique des marques, etc.).

Conçues à l'origine pour fonctionner de manière autonome en assumant toutes les opérations, les installations mécanographiques nous semblent aujourd'hui appelées à jouer une nouveau rôle, très important, à la périphérie des installations électroniques. Dans la mesure où elles peuvent être plus facilement que tout autre équipement disséminées auprès des usagers, à la fois pour élaborer et affiner les données d'entrée sur ordinateur, et pour reprendre et compléter le détail des résultats du gros œuvre traité préalablement sur l'ordinateur et qui leur sont communiqués sous forme de cartes créées automatiquement.

2. *Les ordinateurs* : s'ils ne sont plus « cerveaux », comme naguère, ils sont toujours « électroniques ». C'est-à-dire qu'au lieu de traiter des codes perforés dans des cartes de papier et permettre ainsi le passage (ou non) d'un courant

électrique, ils vont lire, emmagasiner, comparer, compter, etc., des signes transcodés en aimantations microscopiques sur des supports d'un accès de plus en plus aisé et d'une capacité sans cesse accrue qui se lisent à la vitesse de la lumière. Leur succès a entraîné une évolution extrêmement rapide et nous sommes déjà arrivés aujourd'hui à la quatrième « génération » d'ordinateurs alors que nous avons l'impression d'en être toujours à leurs premiers balbutiements. Leurs immenses possibilités sont encore très mal connues car l'élaboration des programmes d'utilisation suit avec un très grand retard le progrès des performances proprement technologiques.

À la différence du dispositif mécanographique esquissé plus haut, dans lequel chaque machine, indépendante des autres, recevait et lisait le fichier complet pour l'exploiter d'une manière particulière (tri, calcul, impression, etc.), l'ordinateur constitue un véritable ensemble, plus ou moins extensible. De ce fait, il peut connaître des configurations particulières (recevoir, par exemple, plus d'unités d'impression, plus de mémoires, etc.), et, par là même, être très différent suivant le cas (vitesse, capacité, langages, etc.).

D'une manière générale, tout ensemble électronique ou ordinateur remplit essentiellement trois fonctions, l'*entrée des informations,* le *traitement,* la *sortie des résultats,* correspondant aux trois parties constitutives. Chacune de ces parties s'est considérablement perfectionnée au cours des quinze dernières années :

a) Les *éléments d'entrée* et de *sortie des données* (ou *périphériques*) comptent aujourd'hui les lecteurs (ou perforateurs) de cartes et de bandes perforées, les lecteurs optiques, les lecteurs de caractères graphiques normalisés, tous les types de bandes et disques magnétiques, les impressions les plus diverses sur tous supports (papier et cartes mécanographiques). Signalons parmi les plus récentes réalisations : des bandes magnétiques où sont enregistrées les données prêtes pour l'édition (mise en page et justification comprise) et pouvant être directement adressées à un atelier de photocomposition pour édition en offset ; l'établissement direct par l'ordinateur de microfiches dont le texte est lui aussi composé automatiquement, ce qui résoud le problème de l'archivage des volumineux listages qui encombraient nos bibliothèques ; possibilité de consultation à distance sur terminal vidéo (unité d'affichage) des informations contenues dans les mémoires, avec modifications et additions à la demande des lecteurs.

b) Les *mémoires* représentent l'équivalent électronique des fichiers, entièrement référencés, dans lesquels toute information est accessible immédiatement (ou presque) par son *adresse.* Il en existe couramment trois sortes : les mémoires sur *bandes* (où les informations, rangées à la suite, attendent surtout un traitement en file ou en série), et sur *disques* ou sur *tambours,* divisés en sections, sur lesquels l'emplacement de chaque donnée est beaucoup plus facilement localisable. En fait, la « mémoire » d'un gros ordinateur est constituée par divers types de *mémoires partielles,* chacune d'elles ayant des performances et des fonctions caractéristiques : sur les *mémoires rapides* vont figurer des données utilisées en permanence pour tel type de travail et qui contient les « registres » et « dictionnaires-machines », par exemple, alors qu'à l'opposé, la *mémoire de masse* pourra

contenir, sur les plus gros modèles, jusqu'à plusieurs centaines de millions de caractères.

c) L'*unité de traitement* est le centre de commande de l'ensemble électronique. C'est d'elle que vont partir deux séries de décisions : 1° celles qui concernent les opérations logiques et arithmétiques ainsi que les programmes enregistrés (opérations de base prévues de manière systématique et laissées dans la mémoire de l'ordinateur) et 2° les séries d'instructions nécessaires pour mener à bien telle suite d'opérations voulues par l'utilisateur sur telle ou telle partie de la machine et que l'on appelle le *programme*. Un *pupitre* est relié à l'unité centrale. Il permet de commander l'ordinateur de l'extérieur par des touches et des boutons et, par l'intermédiaire d'une machine à écrire, d'échanger des messages (ordres, indications d'erreurs, etc.) et de surveiller le travail de l'ensemble.

Je n'insisterai que sur les aspects techniques qui présentent le plus d'intérêt pour les travaux qui nous concernent directement. Il est évident que, par rapport aux possibilités des équipements mécanographiques, les avantages offerts par les ordinateurs sont hors de mesure.

Laissons de côté la vitesse (elle seule pourtant permet d'envisager sans ridicule des travaux sur plusieurs dizaines ou centaines de millions d'unités lexicales) et notons les plus importants :

La possibilité d'enregistrer, de traiter et d'imprimer la totalité des signes graphiques de chaque langue (240 caractères différents peuvent figurer sur la chaîne d'impression des modèles récents) ;

La possibilité (en cours de perfectionnement) d'entrer les informations par lecture directe de textes graphiques, imprimés ou manuscrits, d'où l'élimination de la frappe actuellement encore généralisée (et fort coûteuse) ;

La possibilité de comparer et de classer des éléments de texte de longueur variable (index de mots, de groupes de mots, de séquences, de phrases, etc.) ; la possibilité d'analyser les environnements des mots, possibilité aussi d'élaborer des concordances de toutes sortes, sans aucune limitation du contexte cité;

Par l'effet de programmes *ad hoc* et la consultation de dictionnaires-machines établis à cette fin, la possibilité de rendre plus ou moins automatique la résolution de divers problèmes lexicologiques bien connus (lemmatisation des verbes, analyse des homographies et des polysémies) ;

Possibilité de « dialoguer » avec l'ordinateur sur certains *terminaux* (unités d'affichage à stylet pointeur, à clavier, etc.), c'est-à-dire fournir, à certains moments du travail, des indications complémentaires, non directement programmables parce que trop complexes et qui permettent à l'ordinateur de poursuivre de manière plus fine les opérations systématiques qui lui sont familières.

Je passe sur divers perfectionnements encore plus récents — l'informatique a aussi ses gadgets — pour dire simplement que tous ces moyens sont aujourd'hui assez largement répandus. Il est même courant d'entendre dire, dans les laboratoires qui disposent de ces machines, qu'elles sont toutes loin d'être saturées et que le temps d'utilisation disponible ne manque pas. Alors ?

Eh bien, je crois que le moment est venu de tirer notre premier bilan : au terme de ces quinze années d'utilisation des machines par les lexicologues, force est de conclure qu'à l'exception de quelques tentatives de pointe, jamais systématisées, et souvent sans lendemain, l'ensemble des réalisations est, disons-le, très modeste, compte tenu des possibilités qui s'offraient. Nous tenterons tout à l'heure d'en rechercher la cause.

Envisageons l'inventaire des applications les plus caractéristiques. Elles peuvent se résumer en quelques types de travaux de compilations :

1. *Travaux d'agent de bureau.* Les machines *copient* et recopient dans toutes les présentations voulues et sans ajouter de fautes... s'il en demeure quelques-unes, c'est qu'elles ont été oubliées à l'entrée du dispositif. Les machines *classent* dans des directions multiples les données dont elles disposent.

Ces deux opérations sont à la base des travaux fondamentaux réalisés aujourd'hui pour toute recherche lexicale : les *index des mots* (formes) et les *tables de concordances* (avec contextes). Il n'est pas un centre équipé d'ordinateurs qui n'ait à ce jour tenté d'élaborer pour un linguiste un de ces deux types d'instruments de travail. Ils ont connu bien des perfectionnements :

a) Les *index* se présentent aujourd'hui sous de multiples formes : classements directs, inverses (pour étude des flexions, de la suffixation, des rimes) ; index des mots, de collocations ; index analytiques ou catégoriels, etc.

Les procédés mis en œuvre pour leur élaboration ont été étendus à d'autres documents, comme par exemple les *index bilingues* (ou multilingues) de dictionnaires anciens qui permettent de retrouver la nomenclature de la langue de sortie, jusqu'alors inaccessible [2]. M. de Tollenaere en place ainsi à la fin des volumes qu'il réédite dans sa collection d'anciens dictionnaires flamand-langues vivantes. De mon côté, j'achève un *Index français-latin* des glossaires et lexiques latin-français du Moyen Âge qui permettra de disposer ainsi d'un pseudo-dictionnaire de l'ancienne langue française, alors que nous n'en avons recueilli aucun datant de cette période.

Une utilisation voisine a permis la réalisation des premiers *index chronologiques,* improprement appelés *dictionnaires chronologiques* [3]. Ils consistent à classer, par date d'apparition dans la langue, la liste des divers mots qui ont été jusqu'ici recensés par ordre alphabétique ou par ordre dérivationnel dans les dictionnaires historiques et étymologiques. J'ai moi-même entrepris (et tenu à jour) un tel travail pour le français depuis 1959. Cet *index* d'un type spécial permet non seulement le tri du fichier sur les dates, mais aussi sur toutes les données associées (textes-sources, domaines ou spécialités, etc.).

b) Les *concordances* ont, elles aussi, connu de nombreuses extensions et divers enrichissements : 1° la présence du contexte a entraîné la plupart des auteurs à dégrouper des emplois que le rapprochement formel des vedettes, comme dans les *index,* imposait abusivement. L'analyse des homographes est ainsi aujourd'hui généralisée ; 2° le tri des séquences elles-mêmes a permis une réorganisation des

2. Voir mon étude « L'inventaire des dictionnaires bilingues », *Cahiers de lexicologie,* 2, 1960.
3. Voir par exemple Finkenstaedt, Leise et Wolff, *Chronological English Dictionary,* Heidelberg, 1970.

rubriques : le classement de la concordance ne tient plus compte seulement de l'ordre alphabétique des mots vedettes, mais aussi de celui de leur environnement (collocations) ; 3° dans une perspective comparative, des *concordances différentielles* sont aujourd'hui réalisées en présentant, groupés autour de chaque mot, les contextes correspondants empruntés à diverses traductions dans une ou plusieurs langues, ou à des versions différentes d'un même texte (textes médiévaux, par exemple).

S'inspirant des principes compilatoires qui sont à la base de ces traitements de données, un groupe de chercheurs américains animé par J. Olney de l'U.C.L.A., a eu l'idée d'étendre les possibilités ainsi offertes par les ordinateurs à la réalisation d'une comparaison systématique de toutes les rubriques du *Webster,* le plus important des dictionnaires américains, en traitant parallèlement des éditions développées et des éditions abrégées. Un ensemble de quelques cinquante concordances particulières a été ainsi élaboré ; concordance du vocabulaire des définitions, des exemples, des citations, des données grammaticales, stylistiques, phonétiques, étymologiques, etc. On peut imaginer l'intérêt d'un tel travail pour l'élaboration d'une science critique des dictionnaires et son application à la théorie lexicographique contemporaine.

2. *Travaux de comptable.* Compte tenu des aptitudes des équipements en usage, les disciplines lexicologiques les plus directement favorisées par la mécanisation sont certainement la *lexicométrie* et la *statistique lexicale.* On sait qu'elles sont nées avant l'emploi des machines comptables, mais nous avons tous pu observer l'essor réel qu'elles ont pris plus récemment, tant sur le plan de l'affinement des méthodes que sur celui du nombre et de l'ampleur des applications. Rappelons, par exemple que les premiers travaux sur les vocabulaires fondamentaux, traités manuellement, s'étaient limités à l'inventaire de corpus très modestes, ne dépassant guère 300 000 mots, si l'on accepte la liste de Vander-Becke. Le recours aux moyens mécanographiques, d'abord, a permis d'analyser des corpus doubles ou triples (800 000 à 1 000 000 de mots enregistrés, comme pour l'allemand et l'espagnol fondamentaux que j'ai traités à Besançon, il y a déjà plus de dix ans) ; l'emploi des ordinateurs, ensuite, a permis de réaliser des enquêtes encore plus importantes, sur des corpus pouvant aller jusqu'à 5 000 000 de mots et au-delà (travaux d'Allen à Göteborg, de Zampolli sur l'italien fondamental à Pise). De nombreux *dictionnaires de fréquences — index statistiques,* pour parler plus précisément — récemment élaborés sont aujourd'hui proposés (listes de A. Juilland, et tout dernièrement de Pise, par exemple), dont le plus monumental est sans doute le *Dictionnaire de fréquences du Trésor de la langue française,* établi à partir d'un corpus de 70 000 000 d'occurrences, subdivisé en tranches de 25 années pour les XIX^e^ et XX^e^ siècles, et qui distingue des domaines ou types de langues (prose, poésie, etc.).

Il est aujourd'hui convenu, par un accord tacite, que toute publication d'*index des mots* ou de *table de concordances* doit être accompagnée d'une ou plusieurs listes de données statistiques (par ordre croissant, ou/et décroissant), accompagnées des résultats quantitatifs généraux par catégorie, les divers quotients, etc. On notera aussi en passant que la *lexicométrie* (calculs portant sur des

dénombrements exhaustifs) et même la lexico-statique (calculs de probabilités fondés sur les échantillons d'un corpus) sont encore bien loin d'exploiter les possibilités toutes particulières offertes dans ce domaine par les machines modernes, et auxquelles, il faut bien le dire, il est encore peu demandé en la matière.

3. *Travaux d'archiviste.* C'est ce que nous attribuons aux machines lorsque nous leur confions la tâche de recueillir, classer, conserver, puis de retrouver et de mettre à la disposition des utilisateurs des masses importantes de documents.

Il a été dit plus haut combien ces moyens étaient aptes à enregistrer des données lexicales en respectant toutes les exigences photologiques et à les organiser dans les présentations les plus diverses. Non seulement il est possible de prévoir, à côté de chaque élément enregistré, une place pour des compléments susceptibles d'être ajoutés ultérieurement à tout moment, mais l'accès aux divers niveaux de données peut se faire suivant une infinité de cheminements tant formels que sémantiques ou occasionnels. Leur capacité à enregistrer, traiter et conserver des informations de plus en plus massives s'est elle-même sans cesse accrue. Les fichiers perforés, les bandes perforées et magnétiques, les disques et les tambours constituent, au prix de quelques précautions, des *mémoires indéfectibles,* toujours disponibles et pouvant se dédoubler ou s'associer à volonté.

C'est en se fondant sur de telles performances toujours portées plus loin que les récents projets de *banques de données* se multiplient un peu partout. Les *mémoires de masse,* de capacité illimitée, multidimensionnelles, interconnectables, d'une part, les facilités d'interrogation et d'extraction des données, par lecture directe sur imprimante rapide, terminal avec machine à écrire ou récepteur vidéo, d'autre part, sont les récentes conquêtes technologiques qui en donnent les moyens. Les participants à ce colloque auront l'occasion, s'ils ne l'ont déjà fait, de s'informer directement à la fin de nos journées sur une application de ces possibilités entreprise par la Banque de terminologie de Montréal, une des plus avancée en la matière. Le dispositif qu'elle a réalisé est très simple ; on trouvera, dans la communication de R. Dubuc, la description détaillée du procédé.

Il convient aussi de mentionner que beaucoup de progrès sont à espérer du recours à d'autres techniques documentaires récentes qui ajoutent leurs avantages à ceux que nous connaissons maintenant. Ainsi par exemple, les possibilités nouvelles offertes par les microfiches qui, non seulement peuvent être établies avec l'ordinateur par sensibilisation directe du support à la sortie du système, mais aussi consultées et réexploitées (c'est-à-dire que la microfiche recherchée est identifiée dans la collection de microfiches et présentée sur le terminal vidéo par un système de télévision intégré parallèlement à d'autres informations contenues dans les mémoires magnétiques du système). Une telle association, est encore aujourd'hui très coûteuse : je ne la cite ici que pour illustrer une direction possible de certaines applications particulières, puisqu'elle permet, par exemple, d'obtenir à côté des données textuelles, des informations graphiques (schémas, dessins, etc.) indispensables dans beaucoup de travaux terminologiques.

J'en ai fini avec l'inventaire général de ce qui est aujourd'hui acquis sinon connu. Il serait en effet illusoire de dire que tout ceci est généralisé, mais il serait

aussi faux de penser qu'il ne s'agit que d'expériences exceptionnelles. S'il s'agit bien de réalisations trop peu répandues, elles ne cessent de s'étendre. Elles se caractérisent toutes par le fait qu'elles résolvent essentiellement des tâches *machinales, systématiques* et *massives*. Les machines ne réalisent alors que des actes élémentaires : manipulations, classements, décomptes et identifications, mais qui absorbaient en activités purement matérielles le temps que les chercheurs doivent consacrer à la réflexion. Il est indéniable que, grâce à elles, des travaux ont été entrepris à des échelles irréalisables il y a seulement 20 ans. Mais ces avantages évidents et indiscutables se sont trouvés vite assortis d'inconvénients directement proportionnels, car seule la documentation brute s'est ainsi trouvée stockée en énormes quantités. Et l'on a été rapidement placé devant une situation paradoxale. Pourquoi dépouiller aussi vite autant de textes si par la suite on se trouve submergé par des documents dont le nombre même complique sinon interdit l'analyse ? Les méthodes traditionnelles de dépouillement manuel sélectif n'étaient-elles pas finalement préférables parce que simplement mieux assorties à la capacité de réflexion des chercheurs ?

C'est dans un second temps que les chercheurs, et tout particulièrement les responsables de grands programmes, ont pris conscience que la documentation massive pouvait être ici le plus grand handicap en vue de l'exploitation ultérieure. La solution ne pouvait être imaginée, pour respecter les ordres de grandeur, qu'à partir des machines elles-mêmes. C'est ainsi que, par la force logique des choses, nous sommes entrés depuis 1967 environ dans un nouvel âge de l'informatique lexicale, celle où l'on demande à la machine une participation plus « intelligente ». Parce que les hommes l'ont été pour elle, bien sûr, on peut lui demander de nouvelles fonctions plus élaborées, plus « simulatrices » du travail du chercheur, et surtout associées très étroitement aux multiples décisions qui doivent être prises au cours de tout travail.

Toujours pour simplifier cette présentation, je me bornerai à grouper sous deux rubriques les principales applications en cours ou à l'étude dans cet esprit :

1. *La machine et le choix de l'information lexicale.* L'on est parti ici de considérations très terre à terre qui se sont imposées aux grandes entreprises lexicographiques. En effet, pour faire face à l'impossibilité d'analyser les très nombreux contextes tirés de concordances automatiques lors de la rédaction des articles de dictionnaires, deux solutions ont été imaginées pour en réduire le nombre (en dehors de considérations de « qualité », toujours très délicates à cerner) à partir de décomptes statistiques :

Le Dictionnaire de l'ancien écossais (M. Aitken, Édimbourg) demande à l'ordinateur de ne lui établir qu'une fraction (10%) du nombre total des fiches-concordances pour les mots fréquents (plus de 20 occurrences) ; mais en revanche, tous les contextes des mots plus rares dans le corpus sont exploités ;

Le *Trésor de la langue française* (M. Imbs, Nancy) procède différemment, en s'attachant au double critère de la fréquence des mots et de leur environnement immédiat. Les fiches-contextes qui sont retenues pour l'analyse lexicographique

des mots de haute fréquence sont celles dans lesquelles figurent les collocations (groupes binaires) les plus représentées dans le corpus d'ensemble (6 ou + 6).

D'autres tentatives intéressantes poussent encore plus loin ce dernier principe. Elles visent à mettre en évidence, des contextes « porteurs de sens » parmi l'ensemble des dépouillements plus ou moins significatifs. Ainsi, l'équipe du Centre de lexicologie politique (MM. Wagner et Tournier, ENS de Saint-Cloud) a mis au point un programme d'analyse des co-occurrents contextuels. La machine explore un environnement de longueur variable avant et après le mot pivot, afin de détecter les unités lexicales qui reviennent le plus fréquemment avec lui. Si les regroupements mis en évidence sont surtout thématiques ou notionnels, ils n'en sont pas moins souvent très révélateurs. Là encore, il sera fait appel à des calculs statistiques pour la détection des exemples significatifs à faire figurer dans un dictionnaire qui ne peut — pour des raisons matérielles — tout reproduire.

2. *L'affinage des matériaux bruts.* Les *index* et les *concordances* bruts, fondés sur le traitement automatique des textes sur machines mécanographiques et ordinateurs se sont vite révélés à la fois insuffisants et d'un maniement malcommode alors que seules les formes graphiques étaient prises en compte pour le classement et que les contextes étaient délimités d'une manière fixe. Pour remédier à ces inconvénients, d'autant plus graves que les dépouillements prenaient plus d'ampleur, diverses solutions ont été trouvées :

La lemmatisation automatique permet aujourd'hui, grâce à la réalisation préalable de dictionnaires-machines (dictionnaires de formes fléchies, de verbes, d'homographes, etc.) placés dans la mémoire de l'ordinateur, de procéder automatiquement au regroupement des variantes et des flexions (conjugaisons surtout) sous les formes-vedettes retenues. Divers programmes d'analyse automatique des homographes permettent aujourd'hui de traiter la plupart des cas de manière assez satisfaisante (d'autant qu'elles sont souvent affinées par des analyses complémentaires);

La détermination automatique des contextes variables figure actuellement au programme des applications les plus avancées pour remédier aux insuffisances reconnues des *contextes fixes* (collocations, phrases — d'après la ponctuation — ou nombre de mots fixes, avant et après la forme analysée). Ainsi, par exemple, le Centre dirigé par M. de Tollenaere à Leyde expérimente depuis peu une nouvelle formule associant les critères de délimitation d'après la ponctuation forte et un nombre minimum d'unités lexicales entourant le mot. D'autres songent à s'appuyer sur cette formule en l'enrichissant par divers programmes d'analyse formelle de l'environnement (co-occurrences).

En ce domaine, il faut insister sur le fait que nous ne rencontrons encore aujourd'hui que des applications imparfaites qui, puisqu'elles n'apportent pas toujours de solutions satisfaisantes, indiquent les directions dans lesquelles il reste beaucoup à faire. Ces nouvelles utilisations des possibilités de l'informatique sont encore loin de pouvoir répondre aux besoins réels. Il est vrai que l'utilisation maximale des ordinateurs est bien idéalement l'automatisation de l'analyse, et ici les besoins des lexicologues rejoignent ceux des mécanolinguistes en général.

C'est seulement grâce aux progrès des recherches en analyse formelle des langues, par leurs formalisations pour le traitement automatique que les solutions avancées seront trouvées. Les lexicologues sont à ce titre très conscients de leur solidarité avec les recherches de pointe, mais leurs impératifs spécifiques, en particulier les lourdes charges qui leur incombent au niveau de la réalisation des gros travaux lexicographiques en cours, leur imposent d'imaginer des solutions à court terme.

C'est ainsi qu'à côté d'un axe principal de recherches théoriques qui visent à l'élaboration de *dictionnaires* automatiques inscrits dans la perspective d'une automatisation maximale (totale ?) à laquelle se consacrent certains centres de recherches (M. Zampolli vous donnera tout à l'heure plus de détails à propos de ces travaux), il convient, je crois, d'insister très fortement sur une autre approche, plus pragmatique certes, mais qui a l'immense avantage de nous permettre de tirer dès maintenant le meilleur parti possible des équipements existants. C'est ce que j'appellerai la *lexicologie assistée par ordinateur.*

Alors que l'automatisation complète suppose la résolution préalable de très nombreux et délicats problèmes linguistiques, cette formule (qui évolue d'ailleurs au fur et à mesure des progrès réalisés sur la voie de l'automatisation) tire parti des immenses possibilités du « dialogue » avec l'ordinateur. Le chercheur dispose, pour mener à bien son travail, non seulement d'une information brute soumise aux programmes désormais classiques d'indexation, de tri, de calcul de fréquences, etc., mais aussi de toutes les données complémentaires qu'il peut, grâce à la machine, tirer du matériel précédemment indexé, et surtout des analyses antérieures élaborées par lui et qui sont déjà entrées dans l'ordinateur. C'est ainsi que les phases classiques de l'élaboration d'un dictionnaire pourraient toutes être réalisées avec l'aide de l'ordinateur, bien au-delà, donc, de la phase purement documentaire, comme c'est encore le cas aujourd'hui. L'établissement de la documentation de base (index et concordances) étant tel qu'il est aujourd'hui, les divers choix qui vont conduire à la détermination de la nomenclature, des emplois/sens, demanderaient à bénéficier du secours des multiples possibilités de rapprochement et de comparaison dont peut disposer l'usager d'un ordinateur installé devant un terminal. Il en va de même de toutes les démarches que supposent les opérations de rédaction, classement des sens et définitions, choix des contextes illustratifs, détermination des marques d'usage, stylistique, etc.

Je viens de tenter plusieurs expériences dans ce domaine avec le concours du C.N.U.C.E. de Pise (l'un des organismes de recherches linguistiques sur ordinateur des plus importants d'Europe) qui a bien voulu mettre deux programmeurs à ma disposition cet été. La preuve a été faite qu'avec une programmation peu coûteuse, l'ordinateur peut offrir, à une équipe de trente rédacteurs travaillant simultanément, toutes facilités pour délimiter visuellement, sur un terminal vidéo, les contextes intéressants *avant* l'établissement des concordances. Cela permettrait d'éviter de produire les lourds listages inadaptés à tous besoins sous prétexte d'être multivalents et qui représentent en fait, outre l'incommodité, 40 à 50% de contextes inutiles ou insuffisamment représentés, donc éliminés d'emblée même s'ils pouvaient avoir une importance capitale. De même, installés devant ce terminal, les rédacteurs sont en mesure de demander à la machine

d'organiser, suivant leurs directives, les sens et les acceptations dans l'ordre le plus fin, comme ils peuvent, sur leur clavier, introduire toutes les indications qui permettront ultérieurement de rédiger les définitions.

Si l'on consent à échapper à la contrainte des modèles figés du dictionnaire, il suffit de songer à ce que l'application systématique des programmes de concordance des contextes définitoires cités plus haut pourrait ensuite apporter à une confrontation exhaustive des diverses propositions de définition ainsi élaborées, et les consignes de réécriture qui pourraient en être déduites pour l'élaboration définitive. Et ceci n'est qu'un exemple très fragmentaire. À tous les temps de la réalisation d'un dictionnaire, il est aujourd'hui concevable de faire appel, beaucoup plus que l'on a songé à le faire jusqu'à présent, à des ressources souvent ignorées des lexicographes. Le dictionnaire serait ainsi élaboré *sur* et *avec* la machine, qui, jusqu'à la sortie (automatique elle aussi, par bande magnétique déjà justifiée pour la machine photo-composeuse), assurerait un *contrôle permanent* des diverses opérations, et particulièrement des élaborations rédactionnelles, et de leur synthèse.

Il est ainsi tentant d'imaginer, sans pour autant faire de la science-fiction, ce que pourrait être l'une des applications les plus ambitieuses de l'informatique aux travaux lexicologiques et lexicographiques tenus ici pour inséparables : le *dictionnaire ouvert,* qui serait : 1° un *dépôt cumulatif* de toutes les données lexicologiques connues indexées de manière à permettre les exploitations dans les directions les plus variées ; 2° un *dictionnaire dynamique,* structure lexicographique automatisée (comme en ont imaginé les chercheurs de l'Université de Bonn) pour construire progressivement, et avec une aide sans cesse accrue de l'ordinateur, un dictionnaire du moyen haut-allemand, dans lequel viendront prendre place la totalité des données recueillies et analysées au fil des dépouillements. Cette matière très abondante ne serait pas « éditée » sous les formes conventionnelles que nous connaissons aujourd'hui pour des raisons commerciales évidentes, mais elle serait offerte à la consultation avec les facilités que l'on connaît déjà. Elle permettrait aussi, le moment venu, de réaliser et de diffuser des *répertoires partiels* particulièrement sûrs, dans la mesure où le rêve de tout auteur, qui est de disposer d'un gros dictionnaire pour pouvoir en produire un petit, plus parfait, serait chose acquise.

Je voudrais vous demander de pardonner les généralités auxquelles mon sujet m'a condamné. La rétrospective ouverte que j'ai essayé de dresser pour vous en quelques instants voulait montrer que l'informatique, dont les performances ne cessent de croître, est aujourd'hui de mieux en mieux connue et utilisée dans notre domaine. Pour résumer, je dirai encore que nous avons franchi, depuis cinq ans environ, une seconde étape : la première avait été celle où l'on ne demandait aux machines que des travaux subalternes : copies, classements, calculs simples ; la deuxième fait des machines de véritables collaborateurs techniques de plus en plus qualifiés et de plus en plus proches du chercheur. Il est facile de prévoir qu'avant peu les machines se trouveront au cœur même des opérations de recherche (analyse et mise en forme) et qu'elles seront directement concernées à tous les stades de décision.

Il est indéniable que ces développements ne sont pas encore familiers et que ces observations ne font pas encore partie des idées généralement acceptées. Sans faire le procès de quiconque, ni vouloir anticiper sur l'histoire, je voudrais dire à l'attention de ceux qui, de près ou de loin, sont concernés par ces applications, que de bonnes intentions ne suffisent plus aujourd'hui pour donner toutes leurs chances à des développements que nous souhaitons tous et que tous ont appelés de leurs vœux.

L'entrée dans ce deuxième âge nous impose un minimum de discipline et de responsabilités. Aussi devons-nous tenir compte pour conclure des observations suivantes :

Un trop grand désordre règne dans les travaux effectués, en cours ou en projet. Lorsqu'un chercheur dispose d'un ordinateur dans son université, il peut entreprendre la réalisation (et la publication !) de l'index ou de la concordance brute de tel texte de son choix. Mais ce faisant, il ne participe à aucun progrès. Ces réalisations qui restent en deçà de ce qui peut se faire, qui font souvent double ou triple emploi, n'ont aucun autre mérite que l'information personnelle du réalisateur, non négligeable certes, mais elles sont aussi peu productives que les index manuels réalisés il y a dix ans parallèlement aux index mécanisés ;

De même, il n'est plus possible, dans des perspectives communes, de traiter un texte n'importe comment avec n'importe quoi. Les « sacrifices » et limites imposés aux premiers travaux par les performances réduites des équipements d'alors ne sont plus aujourd'hui admissibles pour de nouvelles recherches. Il importe de confier les travaux spéciaux aux moyens les mieux adaptés, ils existent. Le fait d'accepter des contraintes graphiques, par exemple, interdit d'envisager aujourd'hui les possibilités d'une édition satisfaisante. On doit mettre en garde tous les nouveaux usagers contre le fait que céder à la facilité de répéter plutôt que de renouveler entraîne le risque majeur d'avoir à recommencer sous peu un travail insuffisamment pensé ou notoirement rétrograde.

La solution est bien dans une organisation librement consentie (j'aimerais dire « souhaitée ») de tous ceux qui se consacrent aujourd'hui à de tels travaux. Seule, elle permettra d'envisager une action efficace ; et c'est à cette condition que nous pourrons imposer aux organes financiers de nos universités ou de nos ministères de tutelle, la conviction que nos demandes de crédits sont aussi sérieusement justifiées que celles qui émanent des chercheurs de « sciences exactes ». Et nous savons tous que nous souffrons en premier lieu d'un sérieux handicap en la matière. Cette organisation devrait prendre corps par la mise en place de réseaux informatiques qui en faciliteraient la définition et le fonctionnement. Ceux-ci peuvent se placer à trois niveaux :

a) Les *banques de données et de programmes* disposeraient seules des gros ensembles indispensables pour les traitements massifs : elles trouveraient place à l'échelon national ou international ;

b) Les *centres universitaires ou interuniversitaires de calcul,* dans lesquels une section d'informatique lexicale (ou *linguistique,* numérique ou non) élaborerait les données rassemblées dans chaque centre ou laboratoire de recherche ;

c) *Les équipes de recherche,* dotées du matériel le plus modeste, souvent un seul terminal, chargées de la collecte et de l'interprétation des données. À chacun de ces niveaux correspondraient des tâches bien définies, effectuées sur des équipements spécifiques.

Le premier pas, pour les intéressés, est d'abord de chercher à développer la *coordination* et la *normalisation* entre tous les projets actuellement en cours ; celles-ci passent nécessairement par l'information (ou la formation) réciproque, en prélude à l'élaboration d'un code du travail sur machines.

Cela ne surprendra personne si j'aboutis ainsi à des conclusions voisines de celles des Journées de l'Office de la langue française de Baie Saint-Paul. Il était prévisible que l'informatique appliquée aux travaux lexicologiques (et, par extension directe, aux travaux terminologiques) allait se trouver dans la situation générale des banques de données. Je suis heureux de saisir l'occasion qui m'est offerte pour souhaiter qu'ici encore nos amis du Québec, comme en d'autres domaines, donnent l'exemple. Voilà le bilan promis. Puissent-ils y trouver un argument de plus pour se placer à l'avant-garde des banques de terminologie.

BERNARD QUEMADA