Meta
Journal des traducteurs
Translators' Journal

META

L'automatisation de la recherche lexicologique : état actuel et tendances nouvelles

Antonio Zampolli

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/004637ar
DOI: https://doi.org/10.7202/004637ar

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

Zampolli, A. (1973). L'automatisation de la recherche lexicologique : état actuel et tendances nouvelles. *Meta*, *18*(1-2), 103–138. https://doi.org/10.7202/004637ar

L'automatisation de la recherche lexicologique : état actuel et tendances nouvelles

Cette communication est un complément à celle de B. Quemada. Elle développera quelques-uns des points du panorama qu'il a dessiné, afin d'examiner quelles méthodes et quelles techniques semblent disponibles pour l'application dans un proche avenir de la lexicologie et de la lexicographie assistée par les ordinateurs.

Quelques-unes des méthodologies auxquelles je me référerai sont depuis longtemps connues et utilisées par d'autres secteurs de la linguistique computationnelle (*computational linguistics*) ou de l'information en général, mais elles ne sont pas encore en usage, tout au moins d'un usage courant, dans le secteur de la lexicologie et de la lexicographie. Cette constatation n'implique naturellement pas de jugement négatif sur le rythme de développement de ces disciplines.

Il ne faut pas oublier en effet que les projets et les entreprises lexicologiques et lexicographiques prévoient normalement l'élaboration d'un nombre très élevé de données et que, par conséquent, les instruments et les techniques élaborées par d'autres disciplines exigent une mise au point et un important degré de perfectionnement pour pouvoir être appliquées en lexicologie et en lexicographie.

Les considérations qui suivent sont basées sur les cours et sur les séminaires tenus dans le cadre de la 2e École internationale d'été pour l' « élaboration des données linguistiques et littéraires » qui eut lieu cette année au mois de septembre à Pise. Il s'agit en particulier des cours qui concernent le domaine lexicologique et lexicographique coordonnés par B. Quemada et les expériences de dépouillement effectuées à la Section linguistique du Centre national universitaire de calcul électronique de l'Université de Pise.

Cette section, que je dirige, compte aujourd'hui 20 personnes, et a pour tâche, parmi ses fonctions institutionnelles, celle de mettre à la disposition des humanistes italiens et étrangers, non seulement les ordinateurs électroniques et le personnel technique nécessaire à l'exécution des opérations, mais encore la consultation scientifique et l'étude des procédures.

La section a élaboré une série de programmes pour les dépouillements lexicaux, phonétiques, métriques, statistiques et syntaxiques des textes. Ces programmes sont *généralisés,* en ce qu'ils fonctionnent quels que soient la nature du texte et la langue dans laquelle il est écrit. La section se charge également d'écrire de

nouveaux programmes éventuellement requis pour des élaborations non prévues dans la procédure normale de dépouillement.

Les instituts italiens et étrangers qui effectuent leurs recherches avec la collaboration de la Section linguistique du C.N.U.C.E. sont aujourd'hui plus de 50 et ont enregistré électroniquement plus de 50 millions de mots en 20 langues différentes. Les textes sont enregistrés et élaborés selon les mêmes critères scientifiques et techniques, si bien qu'ils constituent une vaste bibliothèque électronique dans laquelle les textes sont comparables entre eux, et peuvent être élaborés avec les mêmes programmes. La plupart des élaborations sont de nature lexicologique et lexicographique et appartiennent à quatre types fondamentaux : *a*) dépouillements lexicaux pour la rédaction de grands dictionnaires historiques (par exemple : le *Trésor des origines* et le grand *Dictionnaire de l'Accadémie de la Crusca,* les textes historiques de la langue hittite, etc.) ; *b*) dépouillements lexicaux pour la rédaction de dictionnaires historiques synchroniques de langues techniques, ou de disciplines particulières (par exemple : le *Dictionnaire historique de la langue juridique*) ; *c*) dépouillements lexicaux pour la rédaction d'index et de lexiques de *l'œuvre complète* d'un auteur, ou d'un genre littéraire tout entier (par exemple : l'œuvre complète de Sénèque, du philosophe italien A. Rosmini, tout le théâtre latin, etc.) ; *d*) dépouillements lexicaux d'ouvrages (par exemple : Pindare, Bacchylide, les Tabulae Iguvinae, Pétrone, le Code Justinien, la Bible gothique, Goethe, Machado, Vallejo, Baumgarten, etc.).

D'autres élaborations se rapportent aux domaines de la dialectologie (*Atlas linguistique italien*), de la démologie (tous les recueils publiés ou inédits de chants populaires italiens), de la philologie (concordances contrastantes, éditions critiques, comparaison de rédactions successives, etc.), de la psychiatrie et de la psycholinguistique (comparaison entre le langage de malades ou d'arriérés mentaux et le langage « normal » appartenant à des groupes socioculturels analogues), de la *information retrieval* (documentation automatique juridique), de la recherche documentaire (analyse du contenu de documents historiques), de l'histoire de l'art (catalogues de musées ou répertoires archéologiques), de la statistique linguistique (lexiques de fréquence, statistiques phonétiques et syntaxiques de l'italien), de la grammaire (lexique automatique et *parser* syntaxique de l'italien), etc. [1]

1. LE PROBLÈME CENTRAL DE LA LEXICOLOGIE ET DE LA LEXICOGRAPHIE ASSISTÉES PAR LES ORDINATEURS

Les machines mécanographiques ont commencé à se répandre dans les vingt dernières années du siècle passé, mais c'est seulement vers 1950 que P. R. Busa, s.j., mit au point la première procédure mécanographique pour le dépouillement lexical des œuvres de saint Thomas d'Aquin (cf. R. Busa, 1951, et R. Wisbey, 1965). Vers la fin des années 50 les ordinateurs électroniques ont côtoyé et ont peu à peu remplacé les machines UR dans le travail d'élaboration lexicologique et lexicographique, avec plus de dix années de retard par rapport à la diffusion des ordinateurs en Amérique et en Europe. La prise en considération de la nouvelle technologie des ordinateurs ne fut pas sans provoquer quelque perplexité et quel-

1. Pour des informations plus détaillées sur ce projet, voir A. Zampolli, 1973.

ques discussions de la part de nombreux lexicographes et linguistes, habitués à l'emploi des machines UR et ce fut le sujet principal des congrès de Tübingen et de Besançon, respectivement dans l'automne de l'année 1960 et dans l'été de 1961 [2].

Les applications des ordinateurs se divulguèrent par la suite, plus rapidement que ce qui avait été prévu par ces colloques, et le congrès de Prague au printemps de l'année 1966 [3] consacra, pour ainsi dire, d'une manière triomphale, l'avènement des « machines dans la linguistique ». Le panorama des centres pour la mécanisation des recherches lexicologiques existant un peu partout en Europe, la variété des applications et la grandeur des projets exposés suscitaient un grand enthousiasme. Mais vers la fin des années 60 commencent à s'élever contre cet optimisme les premières voix d'insatisfaction et de critique. On se rend compte que le développement des applications suivant l'orientation en quelque sorte classique des années 50 et 60 est arrivé à un point de saturation. Si l'on continue à utiliser la méthodologie courante, selon les actuelles règles du jeu, il n'existe pas de perspectives concrètes de développement.

Les techniques et les méthodologies en vigueur permettent de faire exécuter par l'ordinateur quelques-unes des opérations habituellement réservées au lexicographe ; en particulier, les opérations qui sont liées à la retranscription des contextes pour chaque mot du texte et la mise en ordre alphabétique. Mais, bien que les machines deviennent de plus en plus rapides et les programmes plus sophistiqués, les lexicographes ne peuvent pas en profiter d'une manière proportionnelle car les méthodologies actuelles produisent déjà beaucoup plus de données que ne peut en élaborer une équipe de rédaction de dimension raisonnable, en travaillant selon les procédures actuelles. Nous voici devant une alternative assez claire. Si l'on accepte les limites des applications actuelles, on continuera à enregistrer des *corpus* toujours plus vastes de textes, et l'on constituera des archives de concordances ou de fiches lexicographiques sans aucune analyse ni classification linguistiques préalables. Ces élaborations sont renvoyées à une phase successive de rédaction, sans que l'on se préoccupe pour le moment des énormes problèmes qui résulteront de la quantité de la documentation et des matériaux lexicaux recueillis. Ou bien il faut renouveler les méthodologies actuelles, en appliquant les découvertes de quelques disciplines corrélaires comme la linguistique computationnelle, la statistique linguistique, *l'information retrieval,* et en exploitant d'une manière adéquate les nouvelles possibilités que la technologie des ordinateurs propose dans son rapide développement : en particulier les possibilités d'utiliser un ordinateur comme un instrument de rédaction actif, en interaction étroite avec le lexicographe.

Dans ces affirmations je me suis trouvé en accord avec B. Quemada, dans nos communications au colloque sur *l'élaboration électronique en lexicologie et en lexicographie* organisé par mes soins à Pise dans l'été de l'année 1970 [4]. Notre prise de position fut confirmée au cours de la *1re Table ronde internationale des*

2. Cf. *Cahiers de lexicologie,* 3, 1962.
3. Cf. les actes du colloque : *les Machines dans la linguistique,* Prague, 1968.
4. Cf. A. Zampolli, édit., 1973.

directeurs des entreprises lexicographiques organisée à Florence au printemps de l'année 1971, à l'Académie de la Crusca (cf. de Tollenaere, 1972).

Il fut proposé officiellement d'insérer dans la 2e École internationale d'été de Pise, à côté du programme déjà prévu dédié aux grammaires formelles des langues naturelles et au traitement automatique de structures syntaxiques et sémantiques, une nouvelle orientation pour étudier dans quelle mesure et de quelle façon l'ordinateur peut aider le lexicologue et le lexicographe, aussi bien dans la phase de rédaction et d'analyse des données lexicales que dans leur travail de regroupement [5]. Le développement simultané de ces deux sujets a permis d'observer concrètement combien de thèmes et de problèmes ont en commun la lexicologie et la lexicographie, d'une part, la linguistique computationnelle et de nombreux courants théoriques actuels de la linguistique, d'autre part. Il en a résulté très clairement que l'étude du lexique se présente aujourd'hui comme un thème d'importance centrale également pour ces disciplines. Les linguistes se rendent compte qu'il est indispensable d'affronter le devoir d'appliquer *in extenso* les systèmes de classification et les techniques d'analyse et de description à des sous-ensembles lexicaux toujours plus vastes. Ils sont amenés à réévaluer le travail de regroupement des données, travail qui fut exécuté jusqu'à aujourd'hui presque exclusivement par les lexicographes, et leurs expériences.

Les lexicographes de leur côté, prennent conscience du fait que l'intérêt des linguistes théoriques et des linguistes computationnels, qui converge sur le lexique, a produit des instruments théoriques et techniques qui répondent à certaines exigences fondamentales du travail lexicographique. Ce point de vue a été clairement exprimé également dans les conférences tenues par L. Venezki et par W. P. Lehman à la International Conference on English Lexicography du mois de juin 1972 à New York. Un confirmation ultérieure de cette attitude est constituée par le fait que les applications des ordinateurs à la lexicologie et à la lexicographie feront partie pour la première fois des thèmes de la biennale International Conference on Computational Linguistics, qui aura sa 4e édition à Pise en 1973 [6].

5. L'Ecole internationale d'été, « L'élaboration électronique des données linguistiques et littéraires », a lieu tous les deux ans à Pise. L'édition de 1972 était divisée en deux branches distinctes : *Branche A : Grammaires formelles des langues naturelles et traitement automatique.* Les cours et les conférences se proposaient d'initier les étudiants et les chercheurs aux méthodes de description de la syntaxe des langnes naturelles qui offrent la possibilité d'un traitement automatique. Ils portent essentiellement sur trois arguments : méthodes linguistiques distributionnelles et transformationnelles ; grammaires formelles et automates ; applications des grammaires formelles à l'analyse syntaxique automatique des langues naturelles ; traitement automatique de règles grammaticales. *Branche B : Applications de l'informatique aux analyses lexicales et aux élaborations lexicographiques.* Cet argument a été proposé à l'Ecole par le Convegno internazionale dei Direttori di imprese lessicografiche réuni en mai 1971 à l'Académie de la Crusca. Ces cours et séminaires sont destinés aux étudiants et aux chercheurs qui désirent s'informer sur les possibilités et sur les perspectives de l'application des ordinateurs aux besoins des analyses lexicographiques et de la réalisation des divers types de dictionnaires. Ils porteront essentiellement sur trois matières : problèmes et méthodes de la lexicologie et lexicographie contemporaines assistées par ordinateur ; éléments de statistique lexicale et lexicométrie ; technologie des ordinateurs et des périfériques : leurs applications aux tâches lexicologiques et lexicographiques.

6. Le Congrès se tient tous les deux ans sous l'égide de l'International Committee for Computational Linguistics. En 1965, il eut lieu à New York, en 1967 à Grenoble, en 1969 à Stockholm et en 1971 à Debrecen (Hongrie).

2. SCHÉMA DES DÉPOUILLEMENTS LEXICAUX ACTUELS

Examinons dans ses grandes lignes le schéma d'un dépouillement lexical, de la façon dont il est en général exécuté dans la plupart des centres spécialisés [7].

1re phase : Préparation de l'input. — Le texte est enregistré sur un rapport sur lequel peut travailler l'ordinateur. Les opérations fondamentales sont les suivantes. Le texte est retranscrit par un opérateur, en général sur des fiches mécanographiques. On imprime le contenu des fiches ou de leur équivalent mécanographique, c'est-à-dire que le texte est imprimé par l'ordinateur de la même façon qu'il a été transcrit par l'opérateur. Cette liste est lue et confrontée, en général par le chercheur, avec le texte original, pour rechercher et signaler les inévitables erreurs de recopiage. Ces erreurs sont corrigées à l'aide de fiches. Le texte ainsi corrigé est enregistré définitivement sur une bande magnétique ou sur un support équivalent.

2e phase : Élaborations relatives aux formes graphiques. — L'ordinateur doit en général élaborer une série d'index et de concordances dans lesquelles l'unité d'élaboration est la *forme graphique*. Dans cette phase le mot est donc défini, pour l'ordinateur, comme une séquence de caractères alphabétiques (lettres et signes diacritiques : accents divers, diérèses, apostrophe, etc.) entre deux espaces. L'ordinateur considère deux séquences identiques de caractères alphabétiques comme deux mots égaux, et donc comme deux occurrences d'une même forme graphique. C'est pourquoi, toutes les suites *vers* du texte sont considérées comme des occurrences d'une même force, même si elles ont des sens différents dans les divers endroits du texte (*vers* = préposition, *vers* = unité métrique). L'ordinateur distinguera pour la même raison deux séquences comme *aimer* et *aimons*. Les index les plus communément élaborés sont l'index locorum (liste des formes graphiques suivies chacune par les indications de tous les lieux du texte où elle apparaît), les index alphabétiques direct et inverse des formes et des fréquences relatives, l'index des formes en ordre de fréquence décroissante, divers types de rimes, les concordances des formes et, rarement, les « fiches-contextes [8] ».

3e phase : Lemmatisation. — Sur les listes des concordances par forme, le chercheur analyse linguistiquement les diverses formes. Normalement cette analyse se limite à la lemmatisation. Le chercheur écrit à côté de chaque forme le lemme auquel la forme doit être attribuée. Si la forme est homographe, le lemmatisateur indique le lemme pour chacune de ses occurrences, en examinant le contexte. Un opérateur perfore ensuite ces indications sur des fiches au moyen desquelles l'ordinateur ajoute le lemme à chaque mot du texte enregistré sur bande magnétique. En général, comme on l'a dit, l'analyse ne parvient pas à des niveaux plus

7. Il existe naturellement une grande variété de procédures, mais elles ont en commun une structure unique à laquelle je me réfère.
8. Nous entendons par fiche-contexte une fiche mécanographique qui porte imprimées les données présentes sur la fiche lexicographique traditionnelle (lemme ou exposant, auteur, œuvre, référence, date et un contexte que nous appelons « long » ou « macrocontexte » en opposition avec le contexte « court » ou « microcontexte » des concordances) ; cette fiche porte, aussi, perforées les données essentielles qui permettent d'effectuer automatiquement extractions, insertions, reclassement du fichier (lemme, initiales de l'auteur, œuvre et date).

approfondis. Par exemple elle ne s'effectue pas au niveau morphologique pour distinguer des formes d'un même lemme qui sont homographes sur le plan morphologique (comme *aime*), ni au niveau syntaxique, pour distinguer les différentes constructions ou structures dans lesquelles le mot est introduit, ni au niveau sémantique pour distinguer les diverses acceptions d'un mot polysémique. Pour faciliter et simplifier l'entière procédure de la lemmatisation on utilise quelquefois des écrans terminaux, en particulier si l'analyse linguistique est effectuée aux niveaux énumérés auparavant. À Pise nous avons déjà des procédures de ce type, mais leur usage est encore tellement rare que je préfère en parler par la suite, en tant qu'améliorations à introduire. Souvent, au cours de la lemmatisation, on effectue un choix des matériaux lexicaux ; c'est-à-dire que l'on choisit les lemmes et les occurrences retenues intéressantes pour la phase suivante. Les autres mots reçoivent une marque qui les exclut des opérations successives, par exemple de la production des fiches lexicales.

4e phase : Élaboration des résultats du dépouillement. — On demande à l'ordinateur l'élaboration et l'impression des mêmes index énumérés dans la 2e phase, mais en prenant cette fois pour unité d'élaboration le lemme.

C'est alors que se pose le problème de mettre à la disposition de tous les chercheurs ces résultats. Le fait de les faire recomposer typographiquement présente l'inconvénient de la correction des épreuves. Cet inconvénient est d'autant plus grave si l'on considère qu'en moyenne les résultats du dépouillement consistent en un nombre de lignes qui est environ dix fois celui du texte original.

On préfère les reproduire avec des moyens photographiques ou xérographiques, ou en faisant recours à l'imprimerie au moyen de la photocomposition. Les fiches lexicales sont au contraire produites par ordinateur en deux ou trois séries complètes, qui sont conservées dans des fichiers d'archives.

Toute cette documentation (index, concordances et surtout fiches-contextes) est en général à la base de la rédaction des grands dictionnaires historiques, synchroniques, etc. Il est clair que plus le degré d'analyse est faible au cours du dépouillement, plus sera onéreux et complexe le travail du rédacteur. L'alternative entre l'approfondissement de l'analyse et le choix des matériaux en cours de dépouillement et leur renvoi à la phase de rédaction est aujourd'hui un motif de très vives discussions parmi les lexicologues et les lexicographes qui s'occupent des grands dictionnaires.

3. LES POSSIBILITÉS D'AMÉLIORATION

Au cours de cette conférence je me propose d'indiquer quelques-unes des améliorations possibles pour ce processus. Certaines peuvent être réalisées aussitôt, d'autres au contraire exigent une phase préliminaire de recherche et de travail.

3.1. *La préparation de l'input*

Les difficultés liées à la production des textes en *machine readable form* ont été, et continuent à être, un obstacle important pour l'emploi de l'ordinateur dans le dépouillement et dans l'élaboration des textes [9].

9. Cf. M. Kay, 1967, et A. Zampolli, 1973.

Les difficultés sont soit d'ordre technique, soit d'ordre économique. Il est notoire, en effet, que les claviers des machines ordinairement disponibles pour recopier le texte sur un support lisible par l'ordinateur sont tellement pauvres en caractères que cela engendre une condition difficile et pénible pour représenter la riche variété des caractères et des corps présents dans les textes [10]. Les complications de la codification augmentent encore le nombre des erreurs de frappe. Le chercheur, en essayant de la simplifier, renonce souvent à représenter des graphèmes qui ne lui semblent pas immédiatement indispensables pour les élaborations qu'il a projetées. Il arrive souvent par ailleurs que le texte ainsi appauvri ne soit pas suffisant pour des chercheurs, qui doivent se mettre à le perforer de nouveau. L'insuffisance des caractères disponibles dans les *printer* nuit également à la rapidité et à l'exactitude de la lecture de contrôle.

L'opération de recopiage du texte est en elle-même longue et coûteuse. Le cycle de contrôle et de correction, qui selon notre expérience doit être répété en moyenne 3 fois, triple et le temps et les dépenses. Pour préparer trois pages d'environ 40 lignes chacune, il faut en général trois heures ainsi réparties : perforation une heure, vérification 50 minutes, lecture de contrôle 30 minutes, perforation et exécution des corrections 20 minutes, 2 cycles ultérieurs de contrôle et de correction 20 minutes. Il ne faut pas non plus oublier les temps morts pour le passable des matériaux entre les divers exécuteurs du cycle (contrôleurs, perforateurs, élaborateurs).

Les remèdes à cette situation peuvent être aussi bien techniques qu'organisatifs. L'évolution technologique tend à rendre désuètes les perforatrices des fiches ou des bandes : pour une documentation complète des nouveaux moyens disponibles, je renvoie à l'article de B. R. Scheneider (1971) et j'examine ici uniquement les possibilités offertes par les lecteurs optiques et les terminaux.

Quelqu'un a dit que dans le domaine des lecteurs optiques on est toujours guetté par l'utopie. Tout d'abord on a pensé à un lecteur optique pour un seul *font,* aujourd'hui il existe des lecteurs optiques *multifont,* et déjà on travaille en vue d'un lecteur optique *omnifont.* Ce dernier n'existe pas encore mais on a démontré qu'avec des techniques appropriées il est possible d'ajouter, à peu de frais, de nouveaux fonts à ceux qui sont déjà lisibles par un certain système. Supposons que l'on doive dépouiller un corpus imprimé entièrement avec un seul *font.* On peut calculer s'il convient ou non de faire face aux dépenses de la programmation nécessaire pour ajouter le nouveau *font.* Il est cependant toujours possible de taper le texte avec une machine à écrire pourvue d'un *font* déjà accepté par un système déterminé de lecture optique et de le faire lire ensuite par celui-ci. Il semble être vérifié, par des expériences effectuées exprès, que

10. Il est clair que toutes les informations contenues dans une page imprimée ne sont pas communiquées par les mots et la ponctuation. Les divers corps, la majuscule et la minuscule, le format de la page communiquent divers genres d'information. Par exemple les caractères en italique indiquent des titres, des citations ou des mots étrangers ; il y a des différences qualitatives entre des mots dans des parenthèses ou des crochets ; dans les dictionnaires les relations hiérarchiques sont indiquées par le format et par le type de corps. Toutes les conventions typographiques existent car elles transmettent des informations. C'est pourquoi si le système de codification perd des données typographiques, il perd des informations.

par rapport à la perforation le prix de revient diminue des 3/5 et que le taux d'erreurs est inférieur à 1/5. Le degré de perfection dans la perforation influe par ailleurs sur le bon fonctionnement des contrôles successifs, car les erreurs qui échappent au contrôle sont proportionnelles à la quantité des erreurs présentes. De plus l'enregistrement du texte produit par le lecteur optique peut être une image de la page imprimée. Chaque caractère, chaque corps, et leur place dans la page sont enregistrés automatiquement. Cela signifie qu'il est possible de relever et de marquer automatiquement les catégories d'information indiquées par le type de format, comme par exemple les diverses sections d'un article de dictionnaire.

Un autre système est celui des terminaux CTR (*Cathode Ray Tube Terminal*). Il s'agit d'un terminal qui écrit sur un écran de télévision et tout ce qui apparaît sur l'écran peut être transmis à l'ordinateur. Il y a de nombreux types, parmi lesquels quelques-uns prévoient un écran opaque pour ne pas nuire à la vue, et possèdent un ensemble de caractères très important. Certains types peuvent être même programmés, c'est-à-dire que l'ensemble des caractères est produit par *software,* et peut être multiplié à plaisir.

L'écran de télévision est un instrument dans lequel le texte est parfaitement élastique ; l'effacement est instantané et les erreurs peuvent être localisées, corrigées et vérifiées en une seule opération, au lieu des 6 opérations qui, dans la procédure normale, sont : *a*) imprimer le texte, avec une adresse (numéro d'ordre) univoque pour chaque mot ou ligne ; *b*) noter l'adresse de l'erreur et écrire à côté la correction correspondante ; *c*) perforer l'adresse de chaque erreur, l'ordre opportun (changer en, effacer, ajouter) et les éventuels caractères nouveaux de correction ; *d*) introduire les corrections dans l'ordinateur ; *e*) imprimer les lignes corrigées ; *f*) contrôler que les corrections ont été faites à la place et de la manière voulues.

Avec l'écran, au contraire, l'usager déplace simplement avec une touche le curseur, marque qui indique et détermine le caractère sur lequel on peut opérer à un moment donné. L'usager peut, en utilisant un clavier, effacer, changer, ajouter certains caractères, des mots ou des phrases entières. Si quelque chose est efficace, l'espace vide est rempli automatiquement par le texte qui se déplace vers la gauche ; si quelque chose est ajouté le texte se déplace vers la droite.

Certains terminaux ont aussi la capacité de *formatting,* si bien que la codification dans les applications est simplifiée (par exemple les applications bibliographiques) applications où la même structure est répétée.

Un autre avantage vient du fait que l'ordinateur peut avoir une interaction en ce qui concerne la frappe ou les corrections : le programme opère des contrôles formels (il vérifie par exemple si certaines règles de phonotaxe sont respectées) ou bien lexicaux (par exemple en consultant un lexique automatique)[11].

On peut également faire beaucoup en matière d'organisation nationale et internationale. Dans de nombreux pays certaines maisons d'édition ont adopté

11. Voir les travaux de A. Szanser.

le système d'impression *type-setting* ou la *photocomposition,* et ont ainsi produit une grande quantité de textes sur bande perforée ou sur bande magnétique. En Italie nous sommes déjà en train de travailler en vue d'assurer la disponibilité de ces matériaux pour les recherches linguistiques.

L'objectif le plus important est cependant celui de créer des organismes pour coordonner au niveau national les efforts des divers centres et de chaque chercheur. Il est évident, avant tout, qu'il faut éviter les doubles. Beaucoup de personnes seraient surprises de savoir combien de versions mécanographiques différentes existent pour le même texte, par exemple pour les poèmes d'Homère, la Bible, *la Divine Comédie* et combien de programmes différents exécutent le même type de dépouillement et d'élaboration sur le même type de machine.

Il est nécessaire que les textes soient enregistrés selon les mêmes critères techniques et scientifiques, pour pouvoir être facilement échangés entre chercheurs et pouvoir constituer une grande bibliothèque électronique standardisée, où l'on peut travailler avec les mêmes programmes fondamentaux. L'adoption d'un schéma unique d'enregistrement assure également la reproduction de toutes les informations présentes dans un texte au niveau graphématique, et elle en garantit par conséquent l'emploi pour toutes les recherches possibles. L'adoption d'un tel standard permet non seulement les évidentes économies de temps et d'argent, mais répond également à des exigences d'ordre scientifique : les résultats du dépouillement d'un texte devraient se prêter à la comparaison avec les autres textes du même auteur, de la même époque, de la même école, du même genre littéraire, etc.

3.2. *Utility programs*

La standardisation des archives de textes est une condition nécessaire pour la standardisation des programmes. Il y a aujourd'hui par exemple de très nombreux programmes de contextualisation, c'est-à-dire des programmes qui, pour chaque mot du texte (ou pour certaines catégories de mots du texte), produisent un *record* constitué par au moins trois éléments : le mot, sa référence au texte, son contexte immédiat. Évidemment, les différences entre les divers programmes consistent en l'algorithme qui découpe le contexte.

Après avoir examiné les divers types d'algorithmes en usage dans les centres spécialisés dans le dépouillement des textes, la *Section linguistique* du C.N.U.C.E. a mis au point un programme de contextualisation auquel l'usager peut demander avec un petit nombre de *control-cards,* d'exécuter l'un ou l'autre type d'algorithme [12]. Ce travail de généralisation a été fait pour toutes les phases du dépouil-

12. Les critères de délimitation du contexte s'uniformisent désormais dans les centres de lexicographie qui utilisent des installations mécanographiques. On peut en quelque sorte les regrouper dans les types suivants : *a*) Dans les systèmes de *kwic-index,* mis au point surtout pour des applications documentaires, tous les mots des titres qui composent la liste bibliographique à élaborer, ou plus souvent, les seuls mots lexicaux qui y apparaissent, sont énumérés dans l'ordre alphabétique et reçoivent, comme contextes, les titres ou parties de titres, où ils apparaissent. Les programmes de *kwic-index* aujourd'hui plus diffus, ne sont pas adéquats pour rédiger des concordances dans un but lexicographique, pour plusieurs raisons. *b*) L'exposant est toujours au centre de son contexte, c'est-à-dire qu'il est toujours précédé et suivi par un nombre égal de frappes ou de

lement lexical, du dépouillement phonétique, de quelques élaborations statistiques. Les expériences, exécutées sur environ 2 000 textes pour environ 50 millions de mots en 22 langues différentes, nous permettent d'affirmer que nos procédures sont utilisables, grâce à quelques fiches-contrôles, pour élaborer n'importe quel

---

mots. Le programme est d'une rédaction simple et la composition des contextes très rapide ; souvent, en choisissant cette méthode, on prend la décision de ne pas lemmatiser, afin d'exploiter au maximum la rapidité de la machine et d'éliminer toute intervention humaine. *c*) Le contexte est constitué par une entière unité de référence : le vers, le paragraphe, l'alinéa, etc. *d*) Les limites du contexte sont indiquées en phase de « préédition » : le texte est subdivisé en « péricopes » au moyen de marques qui sont perforées et conservées pendant les élaborations successives : chaque « péricope » fait fonction de contexte pour tous les mots qui le composent. *e*) Le contexte est constitué toujours et seulement par tous les mots compris entre deux signes de ponctuation. Les types *c, d, e,* ont en commun une caractéristique bien précise : le texte est segmenté en « syntagmes » successifs, et tous les mots d'un syntagme ont l'entier syntagme pour contexte, c'est-à-dire qu'ils ont le même contexte ; *f*) Le contexte est choisi selon la nature du mot : pour les mots grammaticaux, c'est souvent un « trinôme » dans lequel le mot grammatical est au centre, pour les prépositions on prend le plus souvent les deux mots suivants, etc. La condition nécessaire est évidemment que les mots qui servent à construire le contexte soient déjà, en quelque sorte, classés ; cette méthode est donc souvent utilisée pour les concordances de lemmes ou de toute façon dans la partie finale du dépouillement (et non comme instrument de travail, par exemple, dans la phase de lemmatisation). A ce sujet sont intéressantes les possibilités d'automatiser, au moins en partie, l'analyse syntaxique, en choisissant pour chaque mot la partie terminale de la structure dont il fait partie et qui est jugée intéressante comme contexte pour la catégorie grammaticale à laquelle appartient le mot. *g*) Le calculateur fournit, avec une méthode quelconque (tout au plus du type *d,* un premier contexte souvent surabondant, qui est ensuite réduit aux dimensions requises au moyen de fiches avec lesquelles on communique à l'ordinateur les mots que le chercheur, après un examen minutieux, a décidé d'éliminer pour alléger le contexte. On peut évidemment rendre cette opération plus simple et rapide en employant un terminal vidéo. *h*) Le contexte est réglé en tenant compte de signes précis comme la ponctuation, le changement de référence, etc. Comme dans le type *b,* le contexte est construit pour chaque mot, si bien qu'il varie d'un mot à l'autre, mais à la différence du type *b,* et à la ressemblance des types *c, d, e,* il est réglé sur la présence d'éléments bien définis, si bien que le mot peut se trouver placé de manière différente dans le contexte : vers le début, vers le centre, vers la fin selon les cas. L'algorithme de notre programme est potentiellement en grade de produire des contextes selon tous les types, même si le type *f* n'a pas été suffisamment expérimenté, du moment que nous utilisons les concordances en phase de lemmatisation avant toute autre analyse. Il est également possible de spécifier quels éléments du texte sont « contextualisés » et ceux qui ne le sont pas, d'énumérer et de classer les éléments qui ont la fonction de « limites » de contexte. Dans notre programme les éléments du texte doivent être classés en : éléments qui doivent faire partie des contextes des autres, mais ne reçoivent pas de contexte propre (par exemple les signes de ponctuation) ; éléments qui doivent avoir un contexte propre et doivent être présents dans les contextes des autres (par exemple les mots « lexicaux ») ; éléments qui ne doivent ni faire partie des contextes, ni recevoir de contexte propre ; en général il s'agit de « codes » introduits dans le texte pour effectuer quelques opérations du programme (par exemple les signes de division en pages et en lignes) ; éléments qui doivent recevoir un contexte propre, mais ne doivent pas entrer dans les contextes des autres (par exemple les mots enclitiques ou les mots qui font partie d'un mot composé, et, en certains types d'élaboration de l'appareil critique, les variants). Avec les éléments du texte on doit également faire une seconde distinction en 2 groupes : éléments qui ont pour fonction de délimiter le contexte (par exemple un signe de ponctuation, ou la fin d'un chapitre) ; éléments qui n'ont pas pour fonction de délimiter le contexte (par exemple un mot). Dans notre algorithme, les limites peuvent être réparties en classes jusqu'à un maximum de 99. Les éléments de la classe 99 sont des « limites insurmontables » de contexte : c'est-à-dire que si en construisant la partie de droite d'un contexte on rencontre une limite de classe 99, le contexte ne se déplacera en aucune façon davantage à droite, quel que soit l'espace encore à sa disposition. On définit souvent comme limites de classe 99 les marques d'un changement de chapitre : parce que naturellement les mots initiaux du chapitre suivant et *vice versa* ne sont pas jugés utiles au contexte du mot final d'un

texte dans n'importe quelle langue, ou tout au moins tous les textes qui peuvent être ramenés à une écriture alphabétique [13].

3.3. *Lemmatisation semi-automatique*

L'ensemble des opérations communément regroupées sous le terme de lemmatisation exige une série d'interventions humaines qui, tout en rompant le rythme des élaborations entièrement automatiques du dépouillement, en augmente considérablement le temps, le coût et les risques d'erreurs. De nombreux chercheurs décident de ne pas lemmatiser du tout les textes, et se limitent à produire des index, des concordances ou des fiches-contextes dans lesquelles les exposants des sont pas des unités définies selon des critères linguistiques, mais de simples *formes graphiques,* c'est-à-dire des mots comme les reconnaît l'ordinateur : séquences de lettres entre deux espaces ou entre deux séparateurs en général.

Sans doute cette simplification a quelque avantage. La rapidité de l'ordinateur qui peut opérer sur des symboles est exploitée complètement et l'on peut produire rapidement, à des prix de revient inférieurs, de grandes quantités de dépouillements qui, une fois divulgués, peuvent rendre d'indéniables services aux chercheurs.

Parfois des raisons scientifiques déconseillent la lemmatisation, par exemple dans le cas de textes en langue ou strates de langue peu connus, dans lequels de nombreux lemmes ne seraient pas attestés ou, vraiment, pas même reconstruisibles.

Toutefois, très souvent, en particulier dans les dépouillements de grands corpus pour la rédaction de vastes dictionnaires historiques, une certaine analyse et classification des matériaux lexicaux semble indispensable avant la conclusion des dépouillements pour éviter que les rédacteurs du dictionnaire soient submergés par la quantité des données à choisir et à ordonner. Il est inévitable de se demander si l'ordinateur, pour aider effectivement le lexicographe, ne doit pas le soutenir aussi et surtout dans la phase de classification des données lexicales que l'ordinateur recueille dans des proportions incommensurables pour les possibilités humaines d'élaboration.

Le pseudo-dictionnaire de machine ou lexique automatique (LA) fournit une première réponse à cette exigence. Pour une plus grande clarté je résume très rapidement ce que l'on entend par LA et consultation d'un LA. Je me référerai constamment, pour faire un plus bref exposé, à la plus simple des diverses structures dans lesquelles un LA peut être organisé (à la note 17 se trouve la description d'une organisation plus complexe).

---

chapitre. Les éléments des autres classes, de la 1re à la 98e, sont des « limites surmontables » de contexte. Si en construisant la partie droite d'un contexte on rencontre une de ces limites, le contexte pourra l'exploiter et se déplacer vers la droite après avoir trouvé une limite de classe égale ou supérieure en se déplaçant dans la direction opposée. Les divers signes de ponctuation peuvent être attribués à diverses catégories, selon leur force : par exemple le point, les points d'interrogation et d'exclamation dans la classe 3 ; les parenthèses, les guillemets, les deux points, le point-virgule dans la classe 2 ; le trait d'union et la virgule dans la classe 1, etc.

13. On doit cependant observer qu'il peut ne pas être convenable de généraliser les programmes de certaines opérations. Je prends comme exemple la consultation d'un lexique automatique très étendu. Dans ce cas, les temps de consultation deviennent un facteur important, dont il est nécessaire de tenir compte. Ils peuvent être améliorés en exploitant quelques caractéristiques morphologiques, propres d'une langue déterminée.

3.3.1. *Le lexique automatique*

Dans sa forme la plus simple un LA consiste en une série de *formes graphiques* enregistrées sur une bande, sur un disque ou sur tout autre support lisible par l'ordinateur. Chaque forme est accompagnée par une série d'informations linguistiques de nature différente selon les divers emplois auxquels le LA est destiné. Nous appelons *fonction* d'une forme l'ensemble des informations qui l'accompagnent.

Si par exemple le LA est rédigé pour aider les dépouillements lexicaux, pour chaque forme seront donnés le lemme auquel la forme doit être attribuée et en général la classification grammaticale et morphologique du lemme et de la forme. Si le LA est rédigé pour traduire automatiquement d'une langue à l'autre on pourra donner également la forme correspondante dans la langue de sortie ainsi que quelques indications pour l'analyse syntaxique et sémantique de la proposition, etc.

Si le LA sert à des statistiques phonématiques, étymologiques, sociolinguistiques, on pourra donner aussi la transcription phonématique de la forme, son étymologie, ses registres d'emploi, etc.

Suivant les buts auxquels le LA est destiné, le nombre des formes qui le composent peut également varier sensiblement. Par exemple si le LA sert à traduire des textes scientifiques relatifs à une discipline spécifique, celui-ci ne contient en général que les mots les plus fréquents dans la langue commune et les termes techniques de la discipline en question.

Si au contraire le LA sert à des statistiques sur le système lexical tout entier, l'ensemble des termes qui composent le LA est beaucoup plus riche. Le LA se propose dans ce cas comme un sous-ensemble représentatif de tout le lexique d'une langue et, à la limite, il voudrait coïncider avec lui [14].

Un LA peut être rédigé avec des processus différents. Dans le dernier exemple cité, on enregistrera tout d'abord un ensemble de lemmes obtenu, en général, par l'union de la nomenclature des principaux dictionnaires imprimés pour une langue donnée. On appliquera ensuite à cet ensemble un algorithme de flexion capable de produire toutes les formes possibles selon le système linguistique donné. Si au contraire on doit lemmatiser un corpus en une langue pour laquelle il n'existe pas encore de dictionnaire valable, on devra procéder graduellement. Les formes du premier texte du corpus seront lemmatisées à la main. Le second texte sera lemmatisé en utilisant un LA qui ne contienne que les formes trouvées dans le premier. Le LA pourra lemmatiser automatiquement seulement les formes du 2e texte déjà apparues dans le premier. Les formes du 2e texte qui sont nouvelles par rapport au premier seront lemmatisées à la main, puis seront insérées dans le LA. Ainsi le LA utilisé pour lemmatiser le 3e texte sera constitué par l'union des formes des deux textes précédents et ainsi de suite.

Du point de vue pratique, il est très important de distinguer les formes univoques des formes homographes. Nous appellerons *univoque* une forme qui,

14. Voir plus avant au paragraphe 3.3.4.

dans le LA, n'a qu'une seule *fonction.* Nous appellerons *homographe* une forme à laquelle, dans le LA, correspondent deux ou plusieurs fonctions distinctes. Par exemple si dans un certain LA la *fonction* des formes est représentée seulement par le lemme, une forme comme l'italien *dica* (dise) sera, dans un tel LA, univoque : cette forme ne peut appartenir en effet qu'au lemme *dire* (dire). Par contre, dans le même LA, la forme *amo* sera homographe, car elle peut aussi bien appartenir au substantif *amo* (ameçon) qu'au verbe *amare* (aimer).

Considérons par contre un autre LA, dans lequel la fonction attribuée à chaque forme comprend, en plus du lemme, également la classification morphologique de la forme (genre, nombre, degré pour les formes nominales ; mode, temps, personne pour les verbes, etc.). Dans un tel LA, sera homographe, non seulement la forme *amo,* mais encore la forme *dica,* auxquelles correspondront 4 fonctions différentes : respectivement 3e personne du singulier de l'impératif, 1re, 2e et 3e personnes du singulier du subjonctif présent. Si au contraire les fonctions d'un LA ne comprenaient que la transcription en alphabet phonétique des formes, dans un tel LA des formes comme *amo* (j'aime) et *dica* (que je dise) seraient univoques, car elles sont dans chaque cas transcrites respectivement en [àmo] et [díka], et ne seraient homographes que les formes qui ne sont pas homophones en italien comme *ancora* ([aŋkora] ancre — [aŋkóra] encore) et *pesca* ([péska] action de pêcher — [péska] de fruit appelé pêche)[15].

3.3.2. *Consultation d'un LA*

L'algorithme de consultation d'un LA ainsi organisé est extrêmement simple. Comme on peut le voir dans l'organigramme de la figure 1, à l'entrée se trouvent le LA (1) et la bande des concordances par formes (2) qui, comme on s'en souviendra, contient les mots du texte dans l'ordre alphabétique, chaque mot enregistré dans un record, accompagné de la référence et du contexte. L'algorithme de consultation lit un record à la fois de la bande 2 et confronte le mot qu'elle contient avec les formes graphiques qui composent le LA. On peut rencontrer 3 conditions différentes :

a) Le mot est identique à une forme qui figure comme univoque dans le LA. — Le programme le lemmatise automatiquement, c'est-à-dire le recopie sur la bande de sortie 5, accompagné, non seulement du contexte et de la référence, mais encore du lemme et des autres éventuelles informations qui sont associées dans le LA à la forme en question.

b) Dans le LA n'apparaît aucune forme identique au mot recherché. — Le programme l'écrit avec les contextes et les références correspondantes sur la bande de sortie 3. À la fin du travail, la bande 3 sera utilisée pour imprimer les concordances de tous les *mots nouveaux,* c'est-à-dire présents dans le texte mais absents dans le LA. Les lemmes (et les autres informations éventuelles) sont écrits à la main dans cette liste, puis sont perforés sur des fiches (6) qui servent soit à lemmatiser les mots nouveaux (bande 10), soit à insérer les nouveaux termes dans le LA (LA intégré avec les formes nouvelles, bande 9).

15. La transcription est donnée, dans le LAI, selon le système phonématique du « florentin cultivé », c'est-à-dire celui que de nombreux auteurs font coïncider avec l'italien standard ou « sans adjectif ».

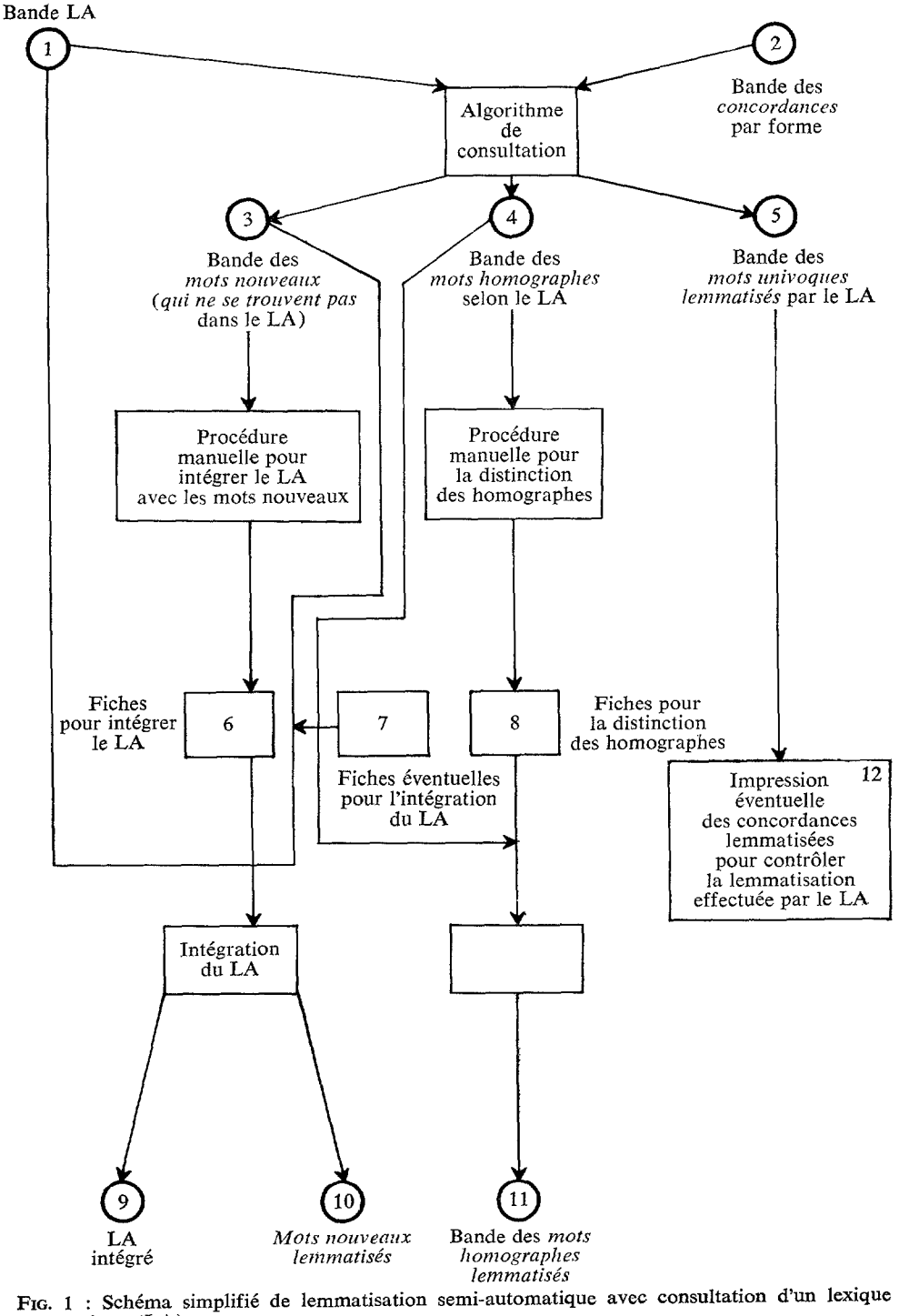

FIG. 1 : Schéma simplifié de lemmatisation semi-automatique avec consultation d'un lexique automatique (LA).

c) Le mot est identique à une forme qui dans le LA figure comme homographe. — Excepté ce que je dirai par la suite sur les possibilités de distinguer automatiquement les homographes, leur lemmatisation doit être faite à la main. Tous les mots homographes possibles [16] du texte sont recopiés sur la bande de sortie 4, accompagnés chacun par le contexte, par la référence et par deux ou plusieurs lemmes (et par d'autres informations éventuelles) proposés par le LA. La bande 4 servira à imprimer les concordances des formes homographes. À côté de chaque forme l'ordinateur imprimera également les lemmes possibles. Le linguiste devra lire les contextes, pour attribuer chaque occurrence de la forme à l'un ou l'autre des lemmes possibles. L'attribution sera ensuite transférée, au moyen de fiches (8) sur la bande de sortie 11. Si l'examen des contextes d'une forme donnée révèle un lemme qui n'est pas compris entre ceux déjà prévus pour telle forme du LA, le lemme doit être ajouté au LA, au moyen de fiches (7)[17].

3.3.3. *Objections à l'emploi d'un LA pour la lemmatisation*

Comme on le sait, les premiers LA ont été mis au point pour la traduction automatique du russe à l'anglais et *vice versa,* à partir de 1946 ; au début, certains considérèrent même qu'un bon LA était une condition nécessaire et suffisante pour traduire automatiquement [18]. Les premiers textes de linguistique computationnelle attribuent une grande place aux systèmes pour rédiger et pour consulter un LA [19]. Toutes ces études ont produit des systèmes très connus et raffinés pour la gestion et la consultation du LA.

On ne doit pas cependant s'étonner si, malgré cela, je place l'emploi des LA parmi les développements possibles de la lexicologie assistée par des ordinateurs. En réalité il y a très peu d'auteurs de dépouillements lexicaux et statistiques qui

16. Nous faisons une distinction entre *homographie possible* dans le système et *homographie actuelle* dans un texte. Il est évident en effet qu'une forme qui, dans le système linguistique considéré, peut appartenir à deux ou plusieurs lemmes distincts, ne réalise en fait, dans un texte donné, qu'un seul de ces lemmes possibles.
17. La méthode de consultation ici décrite exige donc que les mots du texte soient préalablement disposés dans l'ordre alphabétique. Le système de recherche est très simple car il se base sur l'identité graphique absolue entre un mot du texte et une forme correspondante dans le LA. Un système plus complexe, utilisé surtout pour les traductions automatiques et en général quand l'output de la consultation doit être l'input d'un *parser,* est celui de la recherche par thèmes. Dans le LA sont enregistrées non pas toutes les formes graphiques d'un lemme, mais seulement son thème (ou ses divers thèmes ; le mot *thème* ne doit pas être compris dans le sens technique de la linguistique, mais dans le sens de séquence de lettres qui reste invariable au cours de la flexion). Le thème est accompagné d'un code qui renvoie à une table où sont énumérées toutes les désinences possibles, dans le système morphologique considéré, pour la catégorie à laquelle appartient le thème. Mots indéclinables ou formes irrégulières sont enregistrés dans le LA comme des termes indépendants. Si le mot à lemmatiser ne coïncide pas avec un de ces termes, le programme essaie de le segmenter en thème + désinence, de façon à ce que le thème soit présent dans le LA et que la désinence fasse partie de celles qui étaient indiquées comme possibles par la table associée à ce thème. Naturellement quand il y a plus d'une segmentation possible, il s'agira d'un homographe possible. Ce système est plus compliqué mais il permet une considérable réduction des dimensions du dictionnaire. Par exemple, dans le cas de notre LAI, nous avons 150 000 « thèmes » (en comprenant également les mots irréguliers) au lieu de 2 500 000 formes. Pour bien comparer les temps de consultation des deux méthodes, il faut tenir compte de l'entière procédure de dépouillement et des dimensions du texte à lemmatiser. La recherche par *forme,* et non par thèmes, convient quand le texte est exigé dans l'ordre alphabétique.
18. Cf. Mounin, 1964, p. 90ss.
19. Cf. par exemple D. Booth et autres, 1957, Oettinger, 1960, D. Hays, 1967, S. M. Lamb, *in* D. Hays, édit., 1966, et A. F .R. Brown, 1962.

aient adopté un LA pour rendre automatique ou au moins semi-automatique la lemmatisation [20].

3.3.3.1. Nombreux sont ceux qui objectent que la lemmatisation au moyen d'un LA selon la procédure habituelle expose au risque de faire de graves erreurs. La plus grave serait qu'une forme, homographe dans le système linguistique auquel appartient le texte à lemmatiser, soit au contraire insérée, par erreur ou à cause d'une connaissance insuffisante de la langue, comme univoque dans le LA. Dans ce cas l'ordinateur pourrait lemmatiser directement comme univoques les occurrences d'une telle forme dans le texte, sans la soumettre à l'examen du linguiste. Plus le strate de langue que l'on étudie s'étend diachroniquement, plus on recule dans le temps vers des systèmes qui ne sont pas complètement présents à la compétence de celui qui rédige le LA, et plus ce risque est grave.

Toutefois l'obstacle est surmontable si, au lieu de considérer comme définitive la lemmatisation des formes univoques selon le LA, on en imprime les contextes lemmatisés afin que le linguiste contrôle les lemmes attribués par le LA (cf. n.12, fig. 1). Des expériences à ce propos ont montré que l'on pourrait toujours épargner un temps notoire (75% environ) par rapport à une procédure artisanale considérée dans les 3 phases classiques : écrire les lemmes à la main, les perforer et les enregistrer sur bande magnétique, contrôler les 2 opérations précédentes.

3.3.3.2. Une autre objection est celle de ceux qui affirment que l'on devrait créer un LA différent pour chaque dépouillement, car, suivant la nature du texte ou des buts que l'on se propose, l'auteur du dépouillement peut adopter des critères linguistiques différents de lemmatisation.

Cette exigence est en certains cas justifiable, même s'il est évidemment souhaitable que les critères de lemmatisation soient communs pour le plus grand nombre de textes possible : que l'on pense par exemple à la nécessité de pouvoir comparer entre eux les données sur les fréquences lexicales de textes dépouillés par des chercheurs différents.

Toutefois, il n'est pas vrai qu'un LA doive nécessairement produire un seul type de lemmatisation. Dans une certaine mesure, celui-ci permet divers types de lemmatisation, mais en assure en même temps la comparabilité.

Pour illustrer cela voici quelques exemples tirés de l'italien. Nos expériences, et celles des instituts intéressés, montrent en fait que les différences de comportement dans la façon de lemmatiser peuvent être ramenées à deux types principaux. Sous une première rubrique, on peut regrouper toutes les formes dans lesquelles la limite entre homographie et polysémie est incertaine; chaque cas peut être résolu de manière autonome puisqu'il est difficile de les réunir en classes [20]. L'utilisation du LA garantit que chacun de ces cas reçoit le même traitement de

20. Comme l'a écrit Ch. Muller (1963), en l'absence d'un système de règles explicites, le lien sémantique devrait être évalué cas par cas dans la conscience des locuteurs qui, du reste, pourrait être sondée objectivement seulement grâce à des procédés du type « différentiel sémantique » des psychologues. Toutes les études que je connais bien sont d'accord pour affirmer que la seule solution réalisable en pratique consiste à suivre concrètement un dictionnaire, choisi comme arbitre, en tant que représentatif de la communauté des locuteurs. En d'autres termes, en l'absence d'un système de règles, on utilise une liste complète et explicite de toutes les unités.

lemmatisateurs différents. Par contre, l'on place sous la deuxième rubrique les formes qui peuvent être regroupées en classes, une classe comprenant les formes qui présentent toutes la même alternative au lemmatisateur. Par exemple, considérer ou non comme lemmes distincts, respectivement, le masculin et le féminin des noms que l'on appelle mobiles en italien (*maestro-maestra*); les variantes graphiques (*ricuperare-recuperare*) ; l'usage adjectival et substantival (*le amiche-voci amiche*) ; l'usage adverbial et l'usage prépositif (*sopra, sotto, davanti*), etc. Le LA italien (LAI)[21] pourra être adapté au moyen de fiches-contrôles de façon à choisir l'alternative préférée pour chaque classe[22]. L'important est que, grâce

21. La Section linguistique du C.N.U.C.E. a en projet un LAI dont le noyau comprendra un ensemble de 120 000 lemmes environ, obtenus en unissant les lemmes des principaux dictionnaires italiens ; ce noyau pourra être ensuite enrichi périodiquement, surtout par des termes techniques, avec les résultats des dépouillements effectués auprès du C.N.U.C.E. Pour chaque lemme nous enregistrons, dans une première phase d'élaboration, l'appartenance éventuelle à des secteurs particuliers du lexique, les rapports d'homographie ou de « renvoi » à d'autres lemmes, l'étymologie, les affixes et les affixoïdes, la transcription phonématique, une première classification grammaticale, et une série de *codes de flexion* qui ont permis d'obtenir immédiatement la flexion automatique. L'algorithme de flexion a produit les formes, soit en graphie normale, soit en transcription phonématique et nous avons ainsi, à côté d'une liste de lemmes, aussi un dictionnaire de formes ; chaque forme peut être ramenée automatiquement au lemme dont elle provient, et elle est accompagnée aussi par des classificateurs morphologiques traditionnels. Dans une seconde phase, nous enrichirons la série des informations en ajoutant à chaque lemme des classificateurs syntaxiques, sémantiques et statistiques.

22. Nous distinguerons en fait l'homographie dite *radicale* (par exemple *pompa* (magnificence, richesse) et *pompa* (machine à pomper), de l'homographie *fonctionnelle* par laquelle un mot peut être attribué à plusieurs parties du discours (par exemple *sopra* (sur), adverbe et préposition ; *amico* (ami), substantif et adjectif), de l'homographie *morphologique* qui s'établit entre diverses formes d'un même lemme (*dite* = « vous dites » ou « dites », indicatif présent et impératif ; *pensate* = « vous pensez », « pensez » ou « pensées », indicatif, impératif et participe passé). Tandis que des lemmes ou des formes homographiques du deuxième ou troisième type sont suffisamment différenciées par les codes grammaticaux qu'on leur assigne, ceux du premier type exigent souvent une description de nature sémantique. Par exemple *porto* (port de mer), *porto* (porto), *porto franco* (franc de port). En l'absence d'une théorie qui permette d'établir un système de classificateurs sémantiques nous nous limiterons au moins dans un premier temps à distinguer ces lemmes en leur ajoutant une description quelconque ; par exemple quelques synonymes ou une très brève définition. Nous envisageons aussi trois classes différentes de « renvois ». La première classe correspond aux renvois habituels des dictionnaires, du type par exemple de la variante moins courante d'un lemme renvoyée à la variante principale (par exemple *desio* et *disio* « désir »). Grâce à ce type de renvois, l'usager pourra lemmatiser une forme comme *recuperava* (il récupérait) avec un exposant *recuperare* (récupérer), ou bien avec un exposant *ricuperare,* ou bien avec un exposant *ricuperare, recuperare.* Naturellement ce type de renvois se différencie en sous-types particuliers selon que les flexions du lemme variant sont distinctes de toutes les flexions du lemme principal (exemple *recuperare, ricuperare*) ou bien coïncident en partie (exemple *tonare, tuonare* « tonner ») ou bien que le lemme variant soit défectif (exemple *dicere, dire* « dire »). Ce schéma devra être complété, si l'on veut organiser dans le LA des sous-ensembles dialectaux, dont il faudrait indiquer les correspondances de leurs éléments avec les lemmes de l'italien standard. Le second type de « renvois » concerne dans la majorité des cas certaines classes d'homographie fonctionnelle : par exemple, les formes nominales des verbes adjectivées ou substantivées, les adverbes et prépositions, les pronoms et les adjectifs pronominaux, etc. Celui qui utilise le LA pourra considérer *sopra* comme une forme homographe à attribuer respectivement à l'exposant *sopra, adverbe,* ou *sopra, préposition ;* ou bien il pourra le considérer comme une forme univoque et le lemmatiser avec un exposant *sopra, adverbe* et *préposition* ou, plus simplement *sopra.* Les mêmes choix sont possibles pour l'homographie de type morphologique. Ces deux types d'homographie sont très rentables en italien, et leur traitement naturellement influe assez sur le temps global requis par la phase de lemmatisation. La troisième classe de « renvois » concerne essentiellement les unités graphiques qui contiennent plusieurs unités de texte (voir à ce propos A. Zampolli, 1969, et Ch. Muller, 1963) : par exemple, les formes verbales enclitiques, les prépositions articulées, les mots composés

à lui, toutes les formes d'une classe soient traitées suivant le même critère, que ce critère soit formulé de manière explicite, que toutes les formes d'une même classe soient énumérables et qu'il soit possible de rendre facilement homogènes les résultats des dépouillements [23].

3.3.3.3. Une autre objection concerne le coût et le temps requis pour la rédaction d'un LA. En réalité on ne peut guère les négliger : pour rédiger notre LAI du C.N.U.C.E., 7 personnes ont dû travailler pendant 2 ans. Il faut cependant considérer en échange les divers avantages d'un LA. Nous avons déjà parlé de l'épargne de temps et d'argent dans la lemmatisation, épargne qui augmente en proportion du nombre de textes lemmatisés. Le LA assure également une amélioration qualitative dans la lemmatisation. En évitant un grand nombre d'inscriptions de lemmes nécessaires à la procédure artisanale, le LA diminue de manière considérable le risque des erreurs casuelles de transcription et, en même temps, assure une plus grande systématisation et une plus grande cohérence dans les cas, plus nombreux que l'on ne puisse le penser, où la formulation du lemme nécessitait une discussion ou le recours à un corps de règles fixées à l'avance. Le LA fonctionne comme enregistrement des décisions prises et, pouvant être mis à jour et imprimé facilement, fournit à chaque instant le cadre complet des formes déjà examinées et de la façon que l'on a choisie pour les traiter. En outre, le LA permet la rétrodatation automatique des formes dans le corpus choisi.

avec trait d'union, les mots avec affixoïdes. En ce qui concerne les lemmatisateurs les comportements sont souvent différents. On néglige souvent l'enclitique, et une forme comme *parlarmi* (me parler) est attribuée seulement à *parlare*. Mais d'autres fois on constitue aussi une forme *-mi enclitique*, distincte du *mi atone*, qui est attribuée au pronom. Dans ce cas, on se demande aussi si l'on doit conserver sous l'exposant *fare* (faire), les formes *farmi, farti, fargli, farci, farvi, farsi, farlo, farglielo*, etc., ou si l'on doit réunir leurs occurrences en une seule forme *far-*, ou bien simplement les ajouter à la fréquence de *fare*. En ce qui concerne les prépositions articulées on peut décider d'attribuer leurs « occurrences » soit seulement au lemme de la préposition simple, soit seulement au lemme de l'article, soit à l'un et à l'autre. Si on les attribuait aussi à l'article on pourrait : *a*) traiter la préposition articulée comme une réflexion autonome de l'article ; *b*) cumuler les fréquences des différentes préposition articulées en une forme « abstraite » définie comme « forme de l'article lié à des prépositions » ; *c*) ajouter « l'occurrence » des prépositions articulées aux flexions correspondantes de l'article (par exemple à *la* les occurrences de *alla, della, dalla*, à *le* celles de *alle, delle, dalle* etc..). Du reste de semblables alternatives sont possibles pour les formes de la préposition. Dans une phase successive, nous introduirons dans le LA un quatrième type de « renvois », relatifs aux unités de texte composées par plusieurs unités graphiques. Ces unités engendrent des doutes chez le lemmatisateur plus encore pour leur individuation que pour leur traitement ; elles exigent, en plus, que l'algorithme de consultation du LA opère aussi à un niveau syntagmatique. Je me réfère par exemple aux formes des verbes composés avec un auxiliaire et aux soi-disant locutions, que certains ont définies comme groupes de mots qui « préexistent au discours, sont assemblées en langue » (Ch. Muller, 1963). L'énumération explicite et complète des locutions pour l'italien est actuellement, très difficile, que ces locutions soient composées par des mots lexicaux (*segnare il passo* « marquer le pas », *bagnato fradicio* « trempé jusqu'aux os », *avere fame* « avoir faim ») ou par des mots grammaticaux (*al di là* « au-delà », *al di sopra* « au-dessus », *se non altro* « si ce n'est »). Pour s'en rendre compte il suffit de s'en rapporter aux discordances et aux silences des soi-disant bons dictionnaires.

23. En ce qui concerne les effets des critères le lemmatisation sur la validité des dépouillements pour les statistiques lexicales, cf. Ch. Muller, 1963 et 1968, A. Zampolli, 1969, et P. Giraud, 1960. Un exemple d'opération simple pour rendre homogènes les résultats de deux dépouillements menés à partir de critères différents est le suivant. Si les emplois substantivés des adjectifs ont été composés séparément des adjectifs correspondants dans le dépouillement du texte *A*, mais non dans celui du texte *B*, on exclut de la liste des lemmes de *A* tous les adjectifs substantivés, et leurs fréquences s'ajoutent à celles des adjectifs correspondants. Cette opération peut se faire automatiquement grâce au système de « renvois » du LAI.

Le LA favorise généralement la normalisation dans la lemmatisation des textes, et plus le nombre des textes soumis au dépouillement augmente, plus cette normalisation apparaît actuellement indispensable.

3.3.4. *Autres fonctions du lexique automatique*

Le LA, et en conséquence le temps passé à sa rédaction, sont utiles non seulement pour la lemmatisation, mais encore pour de nombreuses recherches linguistiques, parmi lesquelles certaines sont des préliminaires nécessaires pour le développement lui-même de la lexicologie et de la lexicographie, aussi bien en tant que disciplines autonomes, qu'en tant que disciplines étudiées à l'aide de l'ordinateur. Voyons-en quelques exemples.

La linguistique contemporaine est caractérisée par une tendance à formuler les règles exactes et explicites et à énumérer des ensembles complets d'unités linguistiques définis par leurs propriétés. Les ensembles doivent être donnés sous forme de liste d'éléments : « *A system is required to organize the information which has been ascertained, so that this information can be conveniently retrieved when it is required. Such a system is realized as a lexicon* » (H. H. Josselson, 1969). « *I conceive a lexicon as a list of minimally redundant descriptions of syntactic, semantic, and phonological properties of lexical terms* » (Ch. Eillmore, 1968).

On compte déjà quelques exemples de lexique automatique, même établis dans cette perspective [24]. La fonction du lexique est un des arguments centraux même dans les études les plus récentes effectuées dans le domaine des grammaires génératives et transformationnelles. Après que N. Chomsky eut traité cet argument dans *Aspects of Theory of Syntax,* de nombreux chercheurs l'ont développé. L'une des questions principales est celle de l'insertion des éléments lexicaux dans les structures syntaxiques abstraites. Chomsky considère la structure syntaxique comme étant primaire et les éléments qui s'y trouvent insérés comme étant « déterminés par cette structure ». D'autres chercheurs parmi lesquels Lakoff, McCawley, Ross, tendent à considérer comme processus fondamental la construction de la structure sémantique, donc la construction d'une certaine structure lexicale. Je ne peux m'étendre ici sur cet argument. Je suis entièrement d'accord avec M. Gross lorsqu'il affirme à ce propos : « Dans tous les cas, il ne semble pas que des solutions puissent déjà être choisies sur la base des données empiriques que nous possédons. » Un long travail de patience est encore nécessaire pour recueillir des données et l'instrument technique le plus utile est certainement le LA [25].

24. Par exemple pour le hongrois, le LA de L. Papp, 196?, pour le tchèque, le LA de J. Stindlova, 1967 ; Ch. Fillmore a complété une liste de verbes anglais qui entrent, en position *verbe,* dans la construction : *sujet* + *verbe* + *to* + *phrase,* où la phrase introduite par *to* a la fonction de complément du verbe. Dans le LA de H. H. Josselson pour la langue russe, chaque unité lexicale est accompagnée de « *all exsisting phonological, morphological, and syntactic informations* » : en particulier les *constructions* possibles des verbes sont toutes énumérées.

25. En fait M. Gross poursuit une série de recherches sur la représentation des propriétés syntaxiques dans le dictionnaire. Il considère comme indispensable d'effectuer des études systématiques sur un grand nombre d'unités lexicales et sur un nombre élevé de propriétés syntaxiques. Il a recueilli dans le *Lexique des constructions complétives* 1 200 verbes français qui peuvent se construire ainsi. Il a réparti ces verbes en tableaux qui corres-

Le LA permet en outre de mener des recherches statistiques sur le lexique considéré comme faisant partie du système linguistique [26]. Ces recherches ont été jusqu'à nos jours peu nombreuses et insuffisantes non pas parce qu'on les considère moins utiles que celles faites sur les corpus de texte, mais à cause de très grandes difficultés que l'on peut surmonter seulement à l'aide d'un LA.

Les problèmes sont nombreux, qu'ils soient théoriques ou pratiques. Le LA énumère soit les morphèmes [27], soit les mots, et donc nous pourrons effectuer une double statistique tout au moins dans les cas où cela est nécessaire (par exemple le décompte des termes suffixés ne concerne pas évidemment les morphèmes). Mais les définitions du statut linguistique de ces unités sont différentes et contrastantes dans les descriptions structurales de la langue des différents courants. En effet il n'existe pas de définition de telles unités qui soient acceptées universellement et soient entièrement satisfaisantes sur le plan opérationnel [28].

Mais une autre série de problèmes apparaît. Alors que le vocabulaire, c'est-à-dire l'ensemble des mots observés dans un texte ou un corpus, est une liste finie d'unités, le *lexique* d'une langue, lui, est une abstraction, l'ensemble de plusieurs lexiques qui appartiennent à des sous-ensembles fonctionnels.

pondent chacun à une classe distributionnelle de structures. Une étude générale de toutes les formes verbales est en cours au Laboratoire d'automatique documentaire et linguistique du C.N.R.S. : toutes les données sont disponibles sur fiches perforées et sont classées, mises à jour, et imprimées automatiquement. Je pense que nous suivrons cette méthode pour analyser et décrire successivement les sous-ensembles lexicaux de notre LA. Bien qu'ils soient encadrés dans une théorie différente, les articles du LA sur lesquels travaillent A. Zholkovskij et I. A. Melc'huk (1970) sont organisés d'une façon analogue en ce qui concerne la description des diverses « constructions » où chaque mot peut prendre place. Leur LA contient aussi, en plus des descriptions morphologiques et syntaxiques, une *zone sémantique*. Nous aussi, nous nous sommes demandés si nous devions introduire ou non une zone sémantique dans notre LA. Certes le problème de l'adjonction des traits sémantiques aux morphèmes et aux mots est de loin le plus difficile et le plus délicat. La première étape qui consisterait à découvrir méthodiquement de telles propriétés semble être extrêmement complexe. Toutefois, une seule recherche de ce type doit certainement être entreprise en utilisant, au début, des traits classificateurs déjà proposés par différents chercheurs qui se limitent, comme nous le savons, à de petits sous-ensembles du lexique. « *Modern generative grammars of the kind referred to have been more concerned to establish the grammatical classes required in the description of the language they are dealing with than exhaustively to classify all the words in these languages. If all the words of the language are not classified appropriately in the lexicon, the grammar will not be generative in the sens referred as explicit* » (J. Lyons, 1968, p. 159).

26. Entre les textes enregistrés sur bande et le LA, le rapport est dans une certaine mesure analogue au rapport *langue — parole*. Pour les problèmes relatifs au statut du lexique dans la *langue*, voir J. Rey-Debove, 1969. D'autres problèmes sont posés par la formalisation des informations normalement présentes dans un dictionnaire, même s'il faut reconnaître que l'évolution du *métalangage lexicographique* les a rendues toujours plus explicites, au point qu'en grande partie elles peuvent être transcrites directement sur fiche par un opérateur sans préparation formalisatrice préalable.
27. J'emploie ici *morphème* au sens qu'ont donné à ce terme les linguistes américains (cf. F. P. Hamp, 1966, p. 40-41).
28. Un compte rendu critique de la notion de morphème dans la linguistique moderne a été fait par Bierwisch, 1962. Les définitions proposées n'ont pas en général le degré d'exactitude nécessaire pour pouvoir être appliquées à l'analyse d'un grand nombre de données. Comme l'écrit I. I. Revzin, auquel nous devons probablement la première tentative de définition formelle du morphème, le défaut de ces définitions « tient au fait qu'il n'en découle pas de procédure pour segmenter la chaîne parlée en morphèmes » (I. I. Revzin, 1960, p. 98). Toutefois, une définition algorithmique de morphème qui se propose d'être *useful in pratical work in the area of computational linguistics* a été donnée par S. Abraham et F. Kiefer, 1965, p. 9. En ce qui concerne la définition de *mot* la bibliographie est très riche, même si le plus souvent elle concerne la mise en évidence de l'insuffisance et des limites des définitions.

On se pose aussi le problème suivant, souvent débattu mais pas toujours très clairement : le lexique d'une langue doit-il être considéré comme fini ou comme infini [29] ?

Chaque langue peut construire des mots toujours nouveaux par les mécanismes de la dérivation et de la composition. Au niveau du morphème, elle peut introduire de nouvelles unités, par exemple grâce à des emprunts à d'autres langues, pour s'adapter aux nouvelles exigences [30] communicatives de la communauté parlante [31].

Un dictionnaire de lemmes qui, par exemple, résulterait de la nomenclature de tous les dictionnaires d'une langue, pourrait être considéré comme un sous-ensemble représentatif du lexique, dans la mesure où la nomenclature des dictionnaires coïncide avec le vocabulaire d'un corpus représentatif de tous les textes de la langue, ou bien dans la mesure où la compétence lexicale des auteurs des dictionnaires peut être reconnue comme l'équivalent de la compétence de ce « locuteur idéal », auquel se réfèrent les lexicographes qui le considèrent comme un modèle de comportement [32] et les linguistes, spécialement les générativistes, qui le considèrent comme un modèle théorique [33]. On ne peut pas oublier que ce « locuteur idéal » est une abstraction et que la nomenclature d'un dictionnaire correspond à l'union de différents sous-ensembles. Elle doit donc être subdivisée en *zones* distinctes avec des critères les plus explicites possible. Par exemple, dans le LAI, sur l'axe synchronique, on marquera à part « *le parole e le locuzioni toscane, e particolarmente fiorentine, che siano ignote all'uso delle altre parlate italiane* » (B. Migliorini, 1951, p. 13). Il faudra aussi une marque pour les vocables scientifiques et techniques. Ici le problème principal sera la distinction entre les termes utilisés seulement par les spécialistes, et les termes techniques et scientifiques propres aussi au « locuteur idéal », cultivé, mais non spécialiste (J. Dubois, 1971, p. 29). Des marques analogues seront attribuées à d'éventuels mots d'argot, ou d'usage poétique, ou encore à des mots de niveau stylistique particulier, etc. Une partie du lexique restera non marquée, « neutre » pour ainsi dire par rapport aux précédentes catégories ; celle-ci pourra-t-elle être considérée comme

29. Souvent la réponse est différente pour les morphèmes et pour les mots. Voir, par exemple, L. Bloomfield, 1935, chap. 6, L. Hielmsley, 1953, chap. 12 et 14, H. Spang-Hanssen, 1967, R. L. Wagner, 1967. Les statisticiens opposent au *vocabulaire* — un ensemble de mots d'un texte ou d'un corpus, qui est concret, délimité, analysable — le *lexique* « qui [...] est une notion théorique. Aucun Français ne connaît tous les mots en usage de son vivant sur le territoire de la France. Pas davantage n'existe-t-il de dictionnaire qui les enregistre tous sans exception » (R. L. Wagner, 1967, p. 17). « Le lexique est formé d'un ensemble d'unités qui sans être infini au sens mathématique du terme, ne nous donne pas non plus l'impression d'être strictemet fini [...] pour un idiome donné il est impossible de dire quel est le nombre exact des unités qui forment son lexique » (Ch. Muller, 1968, p. 134). Le problème a été aussi examiné sous le jour de la grammaire générative, par exemple par L. Guilbert (1967, p. 116ss.) et par J. Rey-Debove (1970, p. 3ss.) qui différencient nettement la compétence syntaxique de la compétence lexicale ; parce que « si pour la grammaire, la compétence abstraite et la compétence de quelqu'un coïncident *grosso modo* pour le lexique, il y a un abîme entre la première et la seconde ».

30. Cf. Ch. Muller, 1968, p. 138.

31. Cf. L. Hielmslev, 1967, p. 97.

32. Cf. J. Rey-Debove, 1970, p. 7.

33. Cf. Chomsky, 1965, p. 3.

la partie centrale du lexique [34] ? En tout cas le décompte statistique des différentes unités sera exécuté séparément pour le lexique « neutre » et pour chacune des catégories marquées. Naturellement, comme dans toutes les classifications, surtout dans celles qui, en grande partie, sont fondées sur l'intuition du lexicographe, il y aura des zones d'incertitude dans lesquelles le choix est subjectif. Comme le reconnaissent les lexicographes, le seul moyen pour accroître le degré d'objectivité, serait d'avoir recours à la fréquence relative des vocables dans les sous-ensembles d'un corpus représentatif. Les données que nous possédons aujourd'hui étant insuffisantes, il faudra utiliser le LA comme une archive qui s'enrichit perpétuellement : c'est-à-dire y enregistrer les fréquences des unités linguistiques dans les textes qui peu à peu sont dépouillés, et y enregistrer en même temps la position relative des textes par rapport aux axes de la diachronie et de la synchronie (genre littéraire, niveau sociolinguistique, situation régionale, etc.).

Cette considération nous amène encore une fois à réaffirmer la « complémentarité » de la statistique sur les textes et de la statistique sur les dictionnaires, « complémentarité » sur laquelle avaient insisté les représentants de l'École de Prague dès qu'ils ont commencé leurs études sur la statistique phonologique. Je me limiterai à celle-ci pour illustrer les différentes recherches statistiques que nous avons l'intention d'effectuer sur le LAI.

Troubetzkoy lui-même donne des exemples sur la façon de procéder et énumère quelques difficultés de méthodologie [35]. Mais ces difficultés ne suffisent pas à expliquer à elles seules pourquoi le programme des recherches de statistique phonologique tracée par Trobetzkoy et par les premiers pragois a été remarquablement réalisé dans de nombreuses langues [36] en ce qui concerne la fréquence des phonèmes et de leurs combinaisons en corpus de textes considérés comme représentatifs, tandis que les recherches sur le lexique sont beaucoup moins fréquentes et se limitent le plus souvent à des essais méthodologiques sur un nombre restreint de lemmes et de flexions relatives [37].

34. Au stade actuel des travaux, la statistique portant sur les textes offre, pour l'ensemble du lexique, des données dignes de foi seulement pour les mots à grande fréquence. Dans le *Lessico di frequenza dell'italiano contemporaneo* (1971), nous avons retenu 5 365 lemmes (du reste c'est l'ordre de grandeur pour de nombreux dictionnaires de fréquences) sur les 15 000 environ donnés par le dépouillement, et ce dernier chiffre est très inférieur au total des lemmes qui constituent le premier noyau de notre LA, environ 120 000. Les lexicographes se plaignent aujourd'hui de cet état de choses. Par exemple, parlant des limites des applications de la linguistique à la lexicographie, J. Rey-Debove a écrit : « J'ai évoqué l'importance des données de fréquence pour la construction de toute nomenclature, qu'elle comporte 20 000 ou 500 000 mots. Or les données utilisables sur les fréquences ne concernent qu'une petite partie de la nomenclature la plus courte. La statistique lexicale ne vient que préciser l'intuition du lexicographe dans les hautes fréquences et fait justement défaut là où l'intuition ne fonctionne plus » (1969, p. 372).

35. Cf. N. Troubetzkoy, 1939, chap. 7.

36. Les exemples les plus connus sont peut-être les travaux dus à l'école roumaine de linguistique quantitative.

37. N. Saramandu fait allusion de manière explicite à ce chapitre de N. Troubetzkoy dans son travail sur le rendement fonctionnel des types de structures phonématiques. Mais sur les 50 000 lemmes environ du *Dictionarul limbii romane moderne* (Bucarest, 1958) l'on a examiné seulement les 2 903 lemmes formés de 1 à 4 phonèmes et leurs formes flexionnelles ne comportant pas plus de 4 phonèmes (1814). Le travail de S. Golopentia-Eretescu (1966) se rapporte aussi à des matériaux traités par le dictionnaire roumain déjà cité et étudie la structure phonologique des monosyllabes qui y sont rapportées.

Considérons à titre d'exemple l'étude de la « portée fonctionnelle » des différents phonèmes et des oppositions de phonèmes dans une langue donnée, que les pragois placent déjà en 1929 parmi les devoirs fondamentaux de la phonologie synchronique. Appellée peu après *rendement fonctionnel,* elle est définie comme « degré d'utilisation d'une opposition phonologique dans la différenciation des signifiants de mots dans une langue donnée [38] ». L'étude de cette portée fonctionnelle, qui intéressa jusqu'à nos jours de nombreux linguistes [39], nécessiterait en fait la découverte de toutes les *paires minimales,* et leur assemblage en autant de groupes qu'il y a de paires de phonèmes distinctifs [40].

38. Cf. J. Vachek et J. Dubsky, 1960, p. 66.

39. Troubetzkoy, dans le chapitre déjà cité des *Grundzüge,* trace un premier plan de calcul. A. Martinet aussi, par exemple, dans les paragraphes 6.27 et 6.28 de ses *Eléments de linguistique générale* (1968), et dans le paragraphe 3.38, parle de manière explicite de la fréquence des phonèmes dans le lexique. C. F. Hockett, dans *A Manual Phonology* (1935), a défini une méthode pour calculer le rendement fonctionnel d'une opposition sur la base des relations entre le message linguistique et un processus marcovien. C. Tagliavini (1968, p. 117) affirme que la recherche statistique phonologique sur les textes « aurait besoin d'être complétée par un travail mené selon les mêmes paramètres sur un dictionnaire des lemmes et de leurs flexions dans la langue ».

40. Un compte rendu des différentes propositions concernant la définition de *paire minimale* et la bibliographie relative se trouvent dans Z. Muljacic, 1969, p. 200ss. et pour l'italien p. 406ss. L'argument très discuté est étroitement lié aux discussions se rapportant à la « phonématicité », à savoir si l'on doit l'attribuer ou non à un segment phonique. Ici par le terme de « paire minimale irrépréhensible » on désigne *due segni (morfemi o segni maggiori) che contengono gli stessi fonemi tranne uno e nello stesso ordine* (Z. Muljacic, 1969, p. 211). Dans notre LAI il serait possible d'effectuer la recherche des paires minimales des morphèmes. En fait, la plus grande partie des structuralistes travaillent sur les mots et seulement dans quelques cas particuliers, par exemple lorsqu'à l'appui d'une opposition manquent des paires minimales satisfaisantes constituées par des mots, ils recourent à des paires « dans lesquelles l'un des membres, ou les deux, sont des éléments plus ou moins étendus d'un simple mot » (G. Lepschy, 1964, p. 54). Ce type de paire pourra être étudié lorsque nous pourrons appliquer aux textes le LAI et les programmes de recherche syntaxique. L'ordinateur avant tout devrait grouper les formes selon leur longueur. Soit n une de ces longueurs et m le nombre des formes relatives. L'ordinateur devrait produire une première série de m « suites » de longueur n-l obtenues à partir des formes originaires en effaçant le premier phonème de chacune. Après avoir classé alphabétiquement les m suites ainsi obtenues il suffirait de tabuler les suites égales pour trouver les paires minimales qui diffèrent seulement par le premier phonème. En effaçant le deuxième phonème des formes originaires et en classant les m suites (toujours de longueur n-l) ainsi obtenues on pourrait tabuler les paires minimales qui diffèrent seulement par le deuxième phonème, et ainsi de suite jusqu'aux suites dans lesquelles est effacé le $n^{ième}$ phonème. Cette façon de procéder devrait être probablement complétée par d'autres règles, pour tenir compte de certaines observations ayant trait à la définition de paires minimales, citée plus haut. Par exemple, même si désormais il n'est plus jugé nécessaire que les membres des paires minimales appartiennent à la même partie du discours, il est cependant nécessaire que le *contexte phonique* des deux phonèmes en opposition distinctive *soit le même.* Par exemple, l'opposition ne devrait se faire qu'entre segments dont on est sûr qu'ils ont les mêmes traits prosodiques (A. Martinet, 1960, chap. 3.7) : en italien, par exemple, l'accent sur la même syllabe. Il en irait de même pour la structure syllabique et pour les éventuelles jointures ouvertes qui devraient être, dans les deux membres de la paire, semblables (comme on le sait, on a beaucoup parlé de leur valeur pendant les discussions sur le statut phonématique des semi-voyelles et des variétés sonores de *s* et *z* dans l'italien standard : cf., par exemple, R. J. Di Pietro, 1965). Une façon simple pour tenir compte de ces faits, sans modifier l'algorithme précédemment décrit, serait d'insérer dans les suites qui représentent les formes un caractère pour l'accent, un pour la jointure, un pour la séparation entre les syllabes. De cette manière, on ne compterait pas une opposition *e/f* dans la paire *serenata-sfrenata* (sérénade-effrénée) parce que leurs contextes seraient respectivement différents : *sx-re-n'a-ta* et *sxre-n'a-ta* (*x* représente la place du phonème retranché). La situation à ce propos n'est pas très claire. Par exemplde G. Lepschy écrit : « *Di solito non si forniscono coppie minime per l'opposizione di ogni vocale e ogni consonante. Tali coppie sono a rigore impossibili : opporre serenata à sfrenata per fondare una*

Z. Muljacic (1969, p. 307) : « *Per ora è praticamente impossibile trovare tutte le coppie minime in cui possano ricorrere i fenomeni di una lingua, soprattutto i fonemi frequenti. Ciò sara possibile quando i dizionari conterranno tutte le forme di tutti i paradigmi in trascrizione fonologica.* »

En admettant aussi que l'on ait un dictionnaire *imprimé* de ce genre, la découverte des paires minimales ne serait certainement pas facile sur une liste de mots disposés simplement par ordre alphabétique, surtout si l'on tient compte du fait que, dans les statistiques sur le lexique, il n'est pas possible de réduire les quantités des mots à élaborer, comme par contre on peut le faire sur le corpus en utilisant les techniques de l'échantillonnage statistique. Autant que nous sachions, la recherche des paires n'a jamais été faite, même pas pour d'autres langues [41]. Nous croyons qu'il est possible d'essayer seulement avec l'ordinateur, en utilisant un LA de formes.

Les chiffres obtenus grâce à la découverte des paires minimales seront ensuite intégrés à des données sur les fréquences des phonèmes dans les textes, de leur combinaison, et à des données sur les fréquences des mots [42].

Ce type d'intégration est également nécessaire pour les recherches statistiques dans d'autres secteurs. Aussi convient-il d'effectuer une lemmatisation plus « complète » et d'attribuer à la forme, en plus du lemme, des classificateurs qui l'affectent aux différentes catégories linguistiques sur lesquelles porteront les recherches : la provenance étymologique, le mode de dérivation, le système de suffixation, etc.

opposizione /e/:/f/ sarebbe molto artificioso, dato che la struttura sillabica è nei due così diversa e che non si può dire che /e/ ed /f/ si trovino in tali due parole « nello stesso contesto fonico ». Inoltre a quanto pare molte coppie minime per le opposizioni di consonanti e vocali non si troverebbero. Questo punto può essere discusso dal punto di vista della teoria » (G. Lepschy, 1964, p. 61 ; voir aussi Z. Muljacic, 1969, p. 226). En travaillant sur le LA, il sera facile d'effectuer les décomptes suivant les différentes alternatives. En général, l'emploi du LAI permet une grande flexibilité : par exemple compter ou non les occurrences d'oppositions du type *phonème : zéro* (ex. : *dritto-ritto*).

41. « *L'attenzione dei primi linguisti statistici che volevano calcolare il numero assoluto di opposizioni basate su un fonema venima limitata ai fonemi di bassa e bassissima frequenza perché soltanto per essi era lecito sperare di poter stendere un elenco completo di tutte le opposizioni possibili le quali, vista la bassissima frequenza di tali fonemi, non potevano essere numerose* » (Z. Muljacic, 1969, p. 306).

42. Pour l'opposition entre *fréquence d'un phonème dans le lexique* et *fréquence d'un phonème dans le discours,* voir par exemple A. Martinet, 1960, chap. 3.38. Le rendement fonctionnel, comme on sait, est considéré comme élément important pour l'étude des changements phonétiques, en particulier la disparition d'une opposition phonématique, et aussi pour l'étude des mécanismes de reconnaissance des mots et des phrases dans la perception des messages sonores. Tous les linguistes s'accordent à affirmer que la donnée relative au nombre des paires dans le lexique doit être complétée par la donnée de la fréquence d'emploi de leurs membres dans les textes. Différentes techniques pour combiner les deux données ont été proposées. Il est évident que nous nous heurterions à des calculs très complexes et surtout à des problèmes théoriques insurmontables si nous voulions employer un procédé hiérarchique et appliquer ces techniques, non seulement à l'opposition des phonèmes dans le mot mais encore à l'opposition des mots dans la proposition, de la proposition dans la période, etc. Le chapitre 2.4 de *Introduction to Theorical Linguistics* de J. Lyons est probablement l'exposé le plus lucide et le plus complet sur les difficultés d'arriver à une mesure précise du rendement fonctionnel des oppositions paradigmatiques.

Je cite comme exemple à ce propos les recherches de P. Guiraud sur les rapports entre le nombre de phonèmes d'un mot, sa fréquence et ses signifiés [43], et celles de J. Dubois sur la relation entre fréquence et étymologie, et sur la productivité des systèmes de dérivation en français [44].

Le schéma proposé par E. Sapir (1921) pour une classification typologique des langues, est fondé essentiellement sur la segmentation des mots en parties : la racine, les constituants de la dérivation, les flexions. J. H. Greenberg (1960) a proposé une modification de ce schéma qui permet d'utiliser des indices quantitatifs ; par exemple, des rapports comme : n. de morphèmes — n. de mots ; n. de préfixes — n. de mots ; n. de suffixes — n. de mots ; etc. C'est là un exemple d'application statistique rendue possible par l'organisation de notre LA à deux niveaux : morphèmes et mots.

3.3.5. *Tri automatique des homographes*

Une fois que l'on dispose d'un LA et d'un algorithme efficace pour le consulter, pour rendre complètement automatique la lemmatisation il faudrait faire distinguer automatiquement les homographes.

Ce problème a été, et se trouve être central, pour les projets de traduction automatique, tandis que je connais seulement quelques tentatives de la part d'entreprises lexicographiques. Je sais qu'à Nancy l'équipe du *Trésor* avait commencé à étudier des algorithmes pour certains homographes de très haute fréquence, et nous sommes en train de faire la même chose à Pise.

Jusqu'aux années 1966-1967, le problème de la solution automatique des homographes était exposé, en résumé, de la manière suivante.

43. Les premières études de P. Guiraud sur le dictionnaire, par exemple sur la relation entre la fréquence d'emploi des mots et l'étymologie (P. Guiraud, 1954, p. 3), ont été développées par la suite, par exemple, dans le chapitre sur la « structure aléatoire de la dérivation », dans *Structures étymologiques du lexique français,* 1967, où la statistique sur le dictionnaire est utilisée pour confirmer les résultats d' « une analyse purement qualitative et, jusqu'ici, intuitive » qui porte « à imaginer que les signifiés sont des « paquets de sèmes », sèmes qui déterminent leurs relations dans le discours, relations d'où procèdent leurs sens ». P. Guiraud examine par exemple la distribution des mots en relation avec le nombre des signifiés, et la distribution des dérivés sémantiques et morphologiques, en les calculant dans les dictionnaires anglais, français et allemand.

44. Dans l'*Utilisation des statistiques lexicographiques pour l'étude structurale du lexique,* J. Dubois affirme que « la statistique lexicographique fournit sur le plan paradigmatique les données indispensables à l'établissement des modèles du lexique français (proportions relatives des diverses classes syntagmatiques, structures statistiques des dérivés et des composés, extensions relatives des bases) ». Il s'est occupé surtout d'établir le rendement des « composants lexicaux » (affixes) dans des dictionnaires considérés comme représentatifs d'un lexique français défini comme étant « le système le plus large des termes communs d'intercompréhension » (40 000 mots environ). Dans ce but, il calcule par exemple « la proportion des termes suffixés relativement à l'ensemble du lexique » et « la proportion des termes affectés d'un suffixe donné relativement à l'ensemble des termes suffixés d'un microsystème (noms d'action, noms d'agent, verbes, adjectifs) ». Ces proportions permettraient de déterminer le rôle de la suffixation dans la formation du lexique ; elles donneraient une image statistique de l'organisation hiérarchisée de la suffixation, dans chacun des microsystèmes pris séparément, ou bien dans le système général. J. Dubois a étendu ses recherches d'un point de vue diachronique, en consultant les lexiques reproduits dans des dictionnaires de différentes époques. En ce qui concerne les implications relatives à la description formalisée du lexique d'une langue, voir J. Lyons, 1968, p. 159.

Pour l'homographie fonctionnelle, du type substantif-verbe comme l'italien *faccia,* on proposait un *parser* syntaxique [45]. Le principe est évident. Le LA associe à un homographe de ce type non pas une seule description grammaticale mais plusieurs descriptions distinctes, une pour chacune des diverses fonctions syntaxiques que l'homographe pourrait exercer. Si le *parser* réussit et attribue un seul indicateur syntagmatique à la phrase, alors l'ordinateur pourra résoudre automatiquement l'homographie en choisissant la description grammaticale de l'homographe qui a permis au *parser* de réussir.

En ce qui concerne l'homographie de type radical, du type *mozzo della ruota* (essieu de la roue) et *mozzo della nave* (mousse sur un bateau) qui ne comporte pas plusieurs descriptions grammaticales, on proposait un *parser* au niveau sémantique. On pensait par exemple classer tout le lexique présent dans le LA selon des catégories ou, de toute façon, des composantes sémantiques, et de formuler des règles qui spécifient la possibilité ou l'impossibilité de relation, de sélection, de concurrence entre les diverses catégories sémantiques. Dans cette perspective s'inscrivent les diverses tentatives de construire un modèle sémantique global, un réseau de rapports qui relie entre eux tous les mots d'un dictionnaire. Les instruments classiques étaient les divers types de *thesaurus.* Comme chacun le sait, cette façon de procéder a conduit à l'échec des efforts dans la traduction mécanique. Aujourd'hui la situation a changé. Avant toute chose, on tend en général à supprimer, dans les programmes d'analyse automatique, la distinction si nette entre niveau syntaxique et niveau sémantique, parallèlement à la tendance analogue qui s'exprime en linguistique théorique.

Par ailleurs, en partant de la constatation que le langage est un processus de communication entre les gens, et qu'il est inextricablement lié à la connaissance que ces gens ont du monde, certains linguistes computationnels se sont préparés à construire des modèles de cette connaissance, bien que ces modèles soient limités à de petits sous-ensembles bien définis.

Je ne peux évidemment traiter ce sujet, car, comme je l'ai dit, ma conférence veut, de propos délibéré, exclure toute question théorique pour être fidèle au sujet proposé. Je rappelle seulement à titre d'exemple les systèmes de Bobrow et Fraser, Woods, Winograd, et je rapporte le jugement positif qu'en donne P.

45. Les Anglais appellent *parsing,* « décomposition », l'analyse scolaire des phrases. L'analyse d'une phrase peut être effectuée à divers niveaux linguistiques : phonologique, morphologique, syntaxique, etc. Les programmes de *parsing* auxquels nous nous référons, se limitent en général à l'analyse syntaxique et lexicale d'une phrase donnée. Naturellement la description syntaxique qui est l'*output* du *parser* dépend de la grammaire choisie. « *To parse a sentence is to relate it to a general description of a language* » (D. G. Hays, 1966, p. 73). L'input du *parser,* obtenu en général en consultant un lexique automatique, est une séquence d'ensembles de descriptions grammaticales. L'output, comme nous l'avons dit, dépend des diverses théories grammaticales. Par exemple « *a context-free phrase-structure grammar produces strings by rewriting individual symbols; rewriting b as xv in the string abc yields axyc, and rewriting x as wz yields awzyc. The symbol x covers wz, and covers wzy. The substring wz is a constitute of type x, with constituents w and z. To parse a string produced in this manner, it is necessary to find constitutes with constituents in the terminal string (e.g., wz, of type x); then constitutes of which the constituents are either terminal symbols or constitutes already discovered (e.g., xy, of type b); until a constitute is found that covers the whole string* » (Hays, 1966, p. 73-74). Pour un panorama des divers types de *parser* automatiques, voir en plus de Hays, 1966, D. G. Bobrow, 1963, Susumu Kuno, 1966, T. Winograd, 1971, W. A. Woods, 1972.

Lehmann dans le compte rendu du congrès de *computational linguistics* qui eut lieu à Austin (Texas) en 1971 et dans la conférence de New York que j'ai déjà mentionnée, jugement auquel s'associa P. Venezky.

Ces recherches constituent la première réaction positive aux critiques qui, vers le début des années 1960, avaient fait se précipiter la crise de la traduction mécanique. Ces critiques, comme chacun le sait, affirmaient que la traduction mécanique est destinée à l'échec, car il serait impossible de communiquer à l'ordinateur la connaissance du monde et la connaissance de la situation concrète auxquelles se réfère un énoncé, connaissances qui sont une condition nécessaire pour l'analyse et l'interprétation de l'énoncé [46].

Il n'y a pas de doute que la connexion entre les recherches sur les systèmes d'analyse linguistique de l'énoncé et les recherches dans le secteur de l'intelligence artificielle soit très utile, comme l'a affirmé récemment aussi Bar-Hillel, le grand critique de la traduction automatique.

Les systèmes de *parser* utilisés dans les recherches citées ont réussi à traiter correctement des énoncés dont la complexité faisait échouer les *parser* traditionnels, mais il reste le fait que même s'ils traitent une grande variété de structures syntaxiques, ils sont limités à un sous-ensemble lexical très restreint et à des énoncés qui concernent une petite sphère de réalité, dont la connaissance a été formalisée.

L'expérience démontre que la première question que les linguistes posent à ces chercheurs est de savoir si leur système peut être étendu au-delà des limites qu'ils se sont fixées, jusqu'à inclure des sous-ensembles toujours plus vastes et, à la limite, toute une langue. La réponse est toujours la même : la possibilité d'extension du modèle dépend de la possibilité de définir de nouveaux éléments (objets, relation, propriété) et de nouveaux lexèmes au moyen des éléments et des lexèmes précédemment définis avec un appareil formel.

À la différence de la traduction mécanique et des recherches d'intelligence artificielle qui se sont occupées en général de petits sous-ensembles de langue, et le plus souvent de langues techniques, en un certain sens déjà formalisées, les lexicographes soumettent au dépouillement des *corpus* des textes différents, aussi bien dans l'axe synchronique (dialectes, style, situation) que dans l'axe diachronique.

Il pourrait sembler que l'on doive conclure que la solution automatique complète de toutes les homographies est quelque chose d'utopique et que tout effort dans cette direction est destiné à rester improductif.

Malgré cela, il me semble que le lexicographe doit aujourd'hui, plus que jamais, se tenir au courant des développements de la linguistique computationnelle, car les nouveaux *parser* sont beaucoup plus puissants et offrent, dès maintenant, la possibilité de résoudre une bonne partie de l'homographie présente dans les textes qui est produite en grande partie par un petit nombre de mots de très haute fréquence.

46. Voir Susumo Kuno, 1966.

Probablement on devrait faire les premiers pas concrets vers une interaction homme-machine pour résoudre l'homographie, afin d'arriver à un « état de l'art » dans lequel le programme devrait exécuter le *parsing* des phrases pour lesquelles il est adapté, et devrait exiger la collaboration du linguiste pour celles qui dépassent ses capacités. Naturellement on devrait étudier la possibilité d'organiser ce colloque homme-machine en utilisant un écran terminal. Nous avons déjà organisé quelques démonstrations de ce colloque avec B. Quemada et les participants à l'École d'été de Pise en 1972.

Sur l'écran apparaît la forme à analyser. À côté d'elle apparaissent les diverses propositions d'analyses fournies par le LA, numérotées selon l'ordre progressif.

Au-dessous de ces informations, le chercheur peut faire apparaître, l'un après l'autre, les contextes de la forme. Si le *parser* a réussi à analyser quelques-unes des occurrences, le chercheur contrôle l'exactitude de l'analyse choisie. Sinon, il l'exécute lui-même et associe à chaque occurrence, au moyen du clavier ou de la *light-pen,* le nombre progressif qui distingue l'analyse choisie. Au cas où aucune des analyses fournies par le LA ne conviendrait à une certaine occurrence, il ajoutera la nouvelle analyse à celles déjà existantes dans le LA, au moyen du clavier, puis il procédera comme auparavant.

Évidemment ce dernier cas est équivalent, du point de vue de la procédure, au cas de l'analyse d'un texte écrit dans une langue pour laquelle il n'existe pas de LA : au fur et à mesure que l'analyse avance, il se forme un LA qui sera enrichi par les dépouillements successifs.

Dans chaque cas, le chercheur peut demander à l'ordinateur de regrouper immédiatement les contextes qui ont reçu la même analyse, et de faire passer chaque groupe sur l'écran, précédé par l'analyse correspondante et par d'autres commentaires éventuels introduits auparavant. En même temps, il est possible, par exemple, en utilisant la *light-pen,* de faire un choix des contextes à envoyer à l'article de dictionnaire, et éventuellement de « couper » les contextes trop longs de la manière la plus adéquate. Il suffira, en effet, de diriger la *light-pen* sur le premier caractère d'un mot et sur le dernier caractère d'un autre, pour commander à l'ordinateur d'éliminer du contexte tous les mots compris entre les deux points indiqués, en les remplaçant par exemple par des points de suspension entre parenthèses.

En procédant ainsi, non seulement on rend l'analyse plus rapide, mais l'on se dirige également vers la rédaction de l'article de dictionnaire à l'aide de l'ordinateur, afin que se réalisent, dans un même « colloque » avec la machine, la lemmatisation, le choix des exemples, leur « coupe », leur classification en groupes, l'organisation à l'intérieur des groupes, l'insertion des commentaires entre les groupes.

En admettant que le système comprenne une photocompositrice, on pourrait passer directement à l'impression automatique de l'article du dictionnaire.

Un schéma de ce type est certainement et immédiatement réalisable en ce qui concerne la lemmatisation et l'analyse. En ce qui concerne au contraire la rédaction

proprement dite il y a de nombreuses critiques. Une des principales semble être due au fait qu'il ne serait pas possible de réaliser, sur l'écran, l'examen synoptique des divers exemples, chose que le lexicographe est habitué à faire en « éparpillant » les fiches-contextes sur son bureau. On s'est proposé de remédier à cela en mettant deux écrans en parallèle ; sur l'un se trouve fixé le contexte à examiner, tandis que sur l'autre le chercheur fait passer les autres contextes du même mot ; on a pensé également à recourir à des techniques particulières, qui combinent ensemble un écran terminal et les microfiches. Pour une description plus approfondie de ces systèmes, voir Bayley (1973) et la communication de B. Quemada.

3.4. *La banque des mots*

Ce sujet a été traité à fond par B. Quemada et je me limiterai à quelques considérations.

À l'École de Pise se sont dessinées deux interprétations différentes du terme *banque de mots* ou archives lexicales. D'un côté, certains, comme J. Bahr, l'ont entendue comme une sorte de LA, dans lequel chaque mot est accompagné de toutes les informations phonologiques, morphologiques, syntaxiques et sémantiques connues. De l'autre, B. Quemada et moi-même, nous l'entendions comme de véritables archives de mots extraits de corpus de textes analysés, élaborés et non imprimés, mais conservés pour l'ordinateur et mis à la disposition des linguistes et des lexicographes au moyen des techniques d'interaction homme-machine.

Je pense qu'une *banque de mots* ou archives lexicales doive contenir ces deux éléments, le LA et les textes, et que, puisque les archives sont pensées comme dynamiques, aussi bien le LA que le corpus doivent avoir des caractéristiques dynamiques. Le LA représente les connaissances linguistiques actuelles, accumulées par les linguistes et les lexicographes ; ces connaissances représentent la seule structure portante à travers laquelle il est possible d'organiser concrètement, bien que provisoirement, les données du corpus.

Naturellement le LA doit être continuellement modifiable, soit d'après les nouvelles données qui proviennent du corpus, soit à cause de l'évolution des théories linguistiques. Une des informations qui vient du corpus est celle de la fréquence. Ceci donnera progressivement la possibilité de combler une des lacunes justement déplorée par J. Rey-Debove à Cambridge (1969). Pour chaque mot, pour chaque acception, pour les constructions grammaticales auxquels il participe, la banque de mots devrait permettre d'attribuer une fréquence d'emploi, distincte dans les divers sous-ensembles de textes identifiables sur les axes diachroniques et synchroniques [47].

47. En 1949, D. W. Reed, dans un article paru dans *Word*, énonça l'idée que la probabilité d'emploi est un caractère constitutif et non accidentel des unités linguistiques. P. Guiraud, quelques années plus tard, a donné une formulation théorique à ces affirmations. La langue pourrait être comprise comme un ordre systématique d'engrammes, doué chacun d'une probabilité d'emploi, qui constitue un attribut objectif tel que la forme ou le sens. La formulation de G. Herdan est encore plus explicite et radicale. Il propose l'assimilation de l'antinomie saussurienne « langue-parole » d'une part et le rapport « population statistique et échantillon » de l'autre. La « langue » équivaudrait à la totalité des engrammes, dans les cerveaux des membres de la communauté des parlants, chacun avec la respective probabilité d'occurrence. La « parole » équivaudrait « aux échantillons

Naturellement, pour pouvoir organiser le corpus selon la structure d'acceptions, de relations, de propriétés proposées par le LA, il est nécessaire de l'analyser. Cette tâche est d'autant plus ingrate que l'analyse est plus détaillée et le corpus étendu.

Ici se pose à nouveau le problème de savoir comment réduire la disproportion entre les données produites par les dépouillements et les possibilités de l'homme. Nous pouvons seulement énumérer quelques éléments préparatoires pour une réponse provisoire : *a*) L'usage d'un LA fonctionne comme l'enregistrement de toutes les solutions déjà trouvées et suggère le schéma d'analyses. Il indique quels usages, quelles constructions ont été trouvées avec quelle fréquence. *b*) L'usage d'un LA permet un premier niveau d'automatisation de l'analyse. Il permet en premier lieu, comme nous l'avons vu, une lemmatisation semi-automatique. Associé à un *parser* syntaxique, il permet de reconnaître quelques structures. La reconnaissance de certaines structures privilégiées peut exiger un *parser* relativement simple. Prenons l'exemple des diverses constructions d'un verbe. Si dans le LA elles sont indiquées, il est possible de faire regrouper tous les contextes dans lesquels apparaissent des éléments formels, qui indiquent les diverses constructions : par exemple les prépositions.

---

occasionnels de la langue ». Les unités du système linguistique seraient donc caractérisées, non seulement par les traits qualitatifs émergents des oppositions et des relations qui forment la structure du système lui-même, mais encore par les respectives probabilités d'emploi. Ces probabilités ne sont pas directement observables, comme ne l'est pas non plus le système. Elles se traduiraient cependant par le fait, *donné pour certain*, que les unités linguistiques reviendraient dans les textes, parlés et écrits, selon des fréquences relativement stables. Dans cette perspective les fréquences observables dans les textes sont considérées comme des approximations des probabilités non observables dans le système. Beaucoup de personnes qui étudient cette matière reconnaissent encore aujourd'hui, dans cette conception, le fondement théorique sur lequel la statistique linguistique place sa propre autonomie comme science. Ces prémices n'ont pas cependant trouvé aujourd'hui un accord général. Et ceci parce que les résultats des dépouillements de textes, que le développement de la linguistique computationnelle rend toujours plus nombreux et sûrs, démontrent que, à la différence de ce que l'on affirmait, *les fréquences des unités linguistiques dans les textes ne sont pas stables*. Cette constatation semble non seulement rendre contestable le fascinant modèle théorique et global de Herdan et de Guiraud, mais encore les études sur le style de chaque auteur, fondées sur l'espoir de pouvoir donner une mesure objective du style comme écart d'une norme, norme qui perd maintenant consistance et crédibilité. Si ceci est aujourd'hui l'état de l'art, les développements immédiats de la statistique linguistique doivent être destinés à vérifier la structure quantitative de l'univers linguistique. A la suite de Moreau (1961), supposons une définition de l'univers linguistique comme l'ensemble de tous les textes parlés et écrits, produits dans un certain laps de temps. Moreau décrit la fréquence d'un mot dans cet ensemble comme une surface dans un modèle spatial à 3 dimensions : la fréquence, le temps, le genre (livres, revues, lettres, rapports, conversations, etc. forment autant de sous-ensembles plus ou moins homogènes, plus ou moins indépendants). Si la surface qui représente la fréquence d'un mot dans la langue était, dans l'espace à 3 dimensions précédemment défini, un plan horizontal, et si donc la fréquence était la même dans tous les points du plan, le calcul n'exigerait aucune précaution. Mais une telle représentation veut que le mot soit employé de façon identique dans tous les centres d'intérêts et que son emploi ne varie pas au cours du temps. Les dépouillements électroniques ont montré qu'il n'existe pas de conditions aussi absolues. Moreau propose de distinguer les mots en deux groupes : *a*) Les mots de « classe fermée » : c'est-à-dire tous les mots athématiques, ceux qui servent à s'exprimer à propos des choses plutôt qu'à exprimer les choses elles-mêmes. On pourrait placer dans cette catégorie un certain nombre d'adjectifs et de mots courants, de mots grammaticaux, quelques noms très généraux « des termes plus ou moins communs à tous les sujets et à toutes les situations ». Leur emploi varierait dans les divers centres d'intérêts seulement de façon limitée. *b*) Les mots de *classe ouverte* ou « thématiques ». Au contraire, ils présentent des oscillations de fréquence d'un texte à l'autre, et souvent entre les

Naturellement, il resterait des exemples non analysés, qui seraient soumis à l'attention des chercheurs, lesquels interviendraient par exemple au moyen des terminaux. Le LA peut mettre en évidence les lacunes du corpus et suggérer quand il faut recourir aux compétences des informants.

Je désire souligner que l'ordinateur, non seulement fournit la documentation au moyen des dépouillements, mais est également très utile à cause de ses capacités de synthétiser rapidement des structures linguistiques *ad hoc,* dont la « grammaticalité » sera soumise au jugement du parlant, afin d'examiner complètement toutes les possibilités de construction et de classification d'un lemme [48].

Ce dialogue à trois, « LA — corpus — chercheur », me semble le seul système qui nous permette de passer des archives lexicales statiques, imprimées dans une multitude de fiches et de concordances, à une moderne banque des données lexicales qui fasse fructifier les ressources offertes par la technologie moderne. Pour tout cela, il est important évidemment d'avoir une banque de mots centralisée pour chaque langue.

C'est dans ce sens que s'est dirigée l'activité de la Section linguistique du C.N.U.C.E. de Pise. Plus de 2 000 œuvres en cours de dépouillement constituent le premier noyau sur lequel, dès que le LAI sera prêt, nous pourrons essayer de construire la banque de mots de la langue italienne.

ANTONIO ZAMPOLLI

morceaux d'un même texte. Pour évaluer leur fréquence, le seul mode convenable serait de stratifier à priori la langue, de délimiter des centres d'intérêts, à l'intérieur desquels les mots seraient « thématiques ». L'idée de Moreau est justement que les mots d'usage non stable dans la langue, puissent, au contraire, être d'un usage stable dans des textes avec un thème, un centre d'intérêt commun. Il propose pour cela de stratifier une fois pour toutes la langue d'une certaine époque en un certain nombre de centres d'intérêts, qui pourront être également très nombreux, du type par exemple : mathématique, physique, chimie, etc. A l'intérieur de chaque niveau sera déterminée la fréquence des mots thématiques. Par ailleurs la linguistique quantitative a déjà obtenu quelques résultats importants, surtout en ce qui concerne l'application et l'adaptation des méthodes classiques de la statistique aux problèmes de l'échantillonnage lexical. Nous estimons que les études traditionnelles de stylistique et de lexicographie peuvent fournir les données de départ pour une toute première stratification à priori de la langue, ou mieux des données, pour identifier, tout au moins comme hypothèse de travail, quelques sous-ensembles suffisamment définis et également caractérisés par rapport à un paramètre commun. Aux résultats des dépouillements pourront s'appliquer par la suite les techniques statistiques pour vérifier d'une part l'homogénéité des structures quantitatives des sous-groupes et pour identifier, d'autre part, les éventuels mots thématiques. La reconnaissance effective des niveaux de langue devra dériver d'un travail soigné et documenté d'induction à partir de dépouillements toujours plus nombreux, en admettant naturellement que ce travail prouve l'existence, sur le plan des structures quantitatives, de sous-ensembles linguistiques sûrement définissables. Cette attitude, qui se diffuse toujours de plus en plus, est, par ailleurs, favorisée par le développement de la lexicologie automatique dans de nombreux pays où le nombre des dépouillements disponibles s'accroît de jour en jour grâce à l'activité de centres spécialisés. Dans cette même direction les dépouillements et les élaborations pour notre *Lexique de fréquence de l'italien contemporain* constituent une première contribution systématique à l'étude statistique de la langue italienne comme univers lexical. L'élaboration, dans le même but, des données provenant des très nombreux dépouillements qui enrichissent désormais progressivement la bibliothèque électronique du C.N.U.C.E. de Pise, permettra d'associer aux mots du LAI la représentation de leur fréquence subdivisée relativement aux diverses périodes et aux divers genres et styles auxquels appartiennent les textes qui sont peu à peu soumis au dépouillement.

48. Pour plus de détails, voir l'article de R. N. Smith, 1972.

BIBLIOGRAPHIE

Les Machines dans la linguistique. Colloque international sur la mécanisation et l'automation des recherches linguistiques, Prague, 1968.

ABRAHAM, S. et F. KIEFER (1965) : « An Algorithmic Definition of the Morpheme », *Statistical Methods in Linguistics,* vol. IV, p. 4-9.

BAILEY, R. W. (1969) : « Statistics and Style : A Historical Survey », in L. Doležel et R. W. Bailey (édit.), 1969, p. 217-236.

BAR-HILLEL, Y. (1964) : *Language and Information — Selected Essays in their Theory and Application,* Jérusalem.

BIERWISCH, M. (1961) : « Über den theoritischen Status des Morphems », *Studia grammatica,* vol. I, p. 51-89.

BLOOMFIELD, L. (1933) : *Language,* New York.

BOBROW, D. G. (1963) : « Syntatic Analysis of English by Computer. A Survey », in *Proceedings. Fall Joint Computer Conference,* p. 365-387.

BOBROW, D. et J. B. FRASER (1969) : « An Augmented State Transition Network Analysis Procedure », *Proc. of IJCA,* p. 557-568.

BOOTH, D., L. BRANDWOOD et J. P. CLEAVE (1958) : *Mechanical Resolution of Linguistic Problems,* Londres.

BOTHA, R. P. (1968) : *The Function of the Lexicon in Transformational Generative Grammar,* La Haye.

BROWN, A. F. R. (1962) : *The SLC Programming Language and System for MT,* Washington (D.C.).

BUSA, R., s.j. (1951) : *Sancti Thomae Aquinates Hymnorum Ritualium Varia Specimina Concordantiarum. A first example of Word Index automatically compiled and printed with IBM punched card machines,* Milan.

BUSA, R. et A. ZAMPOLLI (1968) : « Centre pour l'automation de l'analyse linguistique (C.A.A.L.), Gallarate », in *les Machines dans la linguistique,* Prague, p. 25-34.

CHOMSKY, N. (1965) : *Aspects of the Theory of Syntax,* Cambridge (Mass.).

DI PIETRO, R. J. (1965) : « The Phonemie Status of Juncture in Italian », in *Proceedings of the Fifth International Congress of Phonetic Sciences,* Basel-New York, p. 261-263.

DUBOIS, J. (1962) : *Etude sur la dérivation suffixale en français moderne et contemporain,* Paris.

DUBOIS, J. (1966) : « Utilisation des statistiques lexicographiques pour l'étude structurale du lexique », in *Statistique et analyse linguistique, Colloque de Strasbourg, 1964,* Paris, p. 95-98.

FILLMORE, Ch. (1968) : « Lexical Entries for Verbs », in *Working Papers in Linguistics,* nº 2, The Ohio State University, p. 23-24.

GOLOPENTIA-ERETESCU, S. (1966) : « La structure phonologique des monosyllabes roumains », *Cahiers de linguistique théorique et appliquée,* vol. III, p. 59-67.

GREENBERG, J. (1957) : *Essays in Linguistics,* Chicago.

GUIRAUD, P. (1960) : *Problèmes et méthodes de la statistique linguistique,* Paris.

HAMP, E. P. (1966) : *A Glossary of American Technical Linguistic Usage,* Utrecht-Anvers.

HAYS, D. G. (1966) : « Parsing », in D. G. Hays, édit., 1966, p. 73-82.

HAYS, D. G., édit. (1966) : *Readings in Automatic Language Processing,* New York.

HAYS, D. G. (1967) : *Introduction to Computational Linguistics,* New York.

HERDAN, G. (1960) : *Type-Token Mathematics,* La Haye.

HJELMSLEV, L. (1953) : *Prolegomena to a Theory of Language,* Bloomington (Ind.).

HJELMSLEV, L. (1957) : « Pour une sémantique structurale », in *Travaux du Cercle linguistique de Copenhague,* vol. XII.

HOCKETT, C. F. (1955) : *A Manual of Phonology,* Bloomington (Ind.).

HYMES, D., édit. (1965) : *The Use of Computers in Anthropology,* La Haye.

JOSSELSON, H. H. (1969) : « The Lexicon : A System of Matrices of Lexical Units and their Properties », in *ICCL*.

KAY, M. (1967) : « Standards for Encoding Data in a Natural Language », *Computers and the Humanities*, vol. I, nº 5, p. 170-177.

KUNO, S. (1966) : « Automatic Syntactic Analysis », in *Seminaron Computational Linguistics*, Bethseda.

LAMB, S. M. et W. H. JACOBSEN, Jr. (1966) : « A High-Speed Large-Capacity Dictionary System », in G. Hays, 1966, p. 51-72.

LEHMANN, W. P. (1972) : *On the Design of a Central Archive for Lexicography in English*, texte préparé pour la International Conference on Lexicography in English, New York.

LEPSCHY, G. C. (1964) : « Note sulla fonematica italiana », *Italia dialettale*, vol. XXVII, p. 53-67.

LYONS, J. (1968) : *Introduction to Theoretical Linguistics*, Cambridge.

MARTINET, A. (1960) : *Eléments de linguistique générale*, Paris.

MEYERSTEIN, R. S. (1970) : *Functional Load*, La Haye.

MIGLIORINI, B. (1951) : *Che cos'è un vocabolario ?*, Florence.

MOREAU, R. (1962) : « Au sujet de l'utilisation de la notion de fréquence en linguistique », *Cahiers de lexicologie*, vol. III, p. 140-158.

MOUNIN, G. (1964) : *la Machine à traduire*, La Haye.

MULJACIC, Z. (1969) : *Fonologia generale e fonologia della lingua italiana*, Bologna.

MULLER, Ch. (1963) : « Le mot, unité de texte et unité de lexique en statistique lexicologique », in *Travaux de linguistique et de littérature*, p. 155-173.

MULLER, Ch. (1968) : *Initiation à la statistique linguistique*, Paris.

OETTINGER, A. G. (1960) : *Automatic Language Translation*, Cambridge (Mass.).

REED, D. W. (1949) : « A Statistical Approach to Quantitative Linguistic Analysis », *Word*, vol. V, p. 235-247.

REVZIN, I. I. (1968) : *les Modèles linguistiques (Modeli jazyka)*, traduit et adapté par Y. Gentilhomme, Paris.

ROSETTI, A. (1947) : *le Mot. Esquisse d'une théorie générale*, Copenhague-Bucarest.

SAPIR, E. (1921) : *Language*, New York.

SCHNEIDER, B. R. (1971) : « The Production of Machine Readible Text : Some of the Variables », *Computers and the Humanities*, vol. VI, n º1, p. 39-47.

SMITH, R. N. (1972) : « Interactive Lexicon Updating », *Computers and the Humanities*, vol. VI, nº 3, p. 137-145.

SPANG-HANSSEN, H. (1956) : « The Study of Gaps between Repetitions », in Morris Halle et al., édit., *For Roman Jakobson*, La Haye, p. 492-502.

SZANSER, A. J. (1969) : *Automatic Error-Correction in Natural Languages*, texte préparé pour *ICCL*, Stockholm.

TAGLIAVINI, C. (1968) : *Applicazioni dei calcolatori elettronici all'analisi e alla statistica linguistica*, in *Atti* del Convegno sul tema : L'automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali (Accademia nazionale dei Lincei, Roma, 1967), Rome, p. 111-118.

TOGEBY, K. (1949) : « Qu'est-ce qu'un mot ? », in *Travaux du Cercle linguistique de Copenhague*, vol. V, p. 97-112.

TRUBETZKOY, N. S. (1939) : « Grundzüge der Phonologie », in *Travaux du Cercle linguistique de Prague*, vol. VII.

VACHEK, J. et J. DUBSKY (1960) : *Dictionnaire de linguistique de l'Ecole de Prague*, Utrecht-Anvers.

VENEZKI, L. (1972) : *Computer Applications in Lexicography*, texte préparé pour la International Conference on Lexicography in English, New York.

WAGNER, R. L. (1967) : *les Vocabulaires français*, Paris.

WINOGRAD, T. (1971) : *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*, Cambridge (Mass.).

WISBEY, R. (1965) : « Computers and Lexicography », in D. Hymes, 1965.

WOODS, W. A. et R. M. KAPLAN (1971) : *The Lunar Sciences Natural Language Information System,* BBN, Cambridge (Mass.).

WOODS, W. A. (1972) : *An Experimental Parsing System for Transition Network Grammars,* BBN Report N. 2362, Cambridge (Mass.).

ZAMPOLLI, A. (1969) : *Appunti per l'intervento alle giornate di studio sul tema « La preparazione del Personale per la elaborazione automatica dei dati, in Italia » (AICA),* Rome.

ZAMPOLLI, A. (1973) : « La section linguistique du C.N.U.C.E. », in A. Zampolli, édit., 1973, p. 133-199.

ZAMPOLLI, A., (édit. (1973) : *Elaboration électronique des données linguistiques et littéraires. Actes de la I^re Ecole d'été internationale de Pise et du colloque « L'élaboration électronique en lexicologie et lexicographie » (Pise, 1970),* Florence.

ZAMPOLLI, A., édit. : *Elaboration électronique de textes et de structures,* Florence (sous presse).

ZHOLKOVSKIJ, A. K. et I. A. MEL'CHUK (1970) : « Sur la synthèse sémantique » *T.A. Informations,* vol. II, p. 1-85.

QUESTIONS

M. Baldinger : J'aimerais attirer l'attention sur un problème qui a été frôlé sans qu'on l'ait discuté dans les deux exposés. Je rapproche la remarque que M. Zampolli vient de faire sur l'exhaustivité et ce que M. Quemada a dit sur le manque de temps pour lire 80% des exemples en stock du T.L.F., par exemple. Il y a presque une quinzaine d'années, j'ai fait un travail sur les dénominations du paysan libre en France. J'ai lu 4 000 documents, je les ai dépouillés. Cela me donnait un résultat scientifique satisfaisant, des données claires, des résultats clairs. Mais à ce moment un médiéviste de l'Université de Bâle me disait : « Qu'est-ce que vous voulez, ces 4 000 documents que vous avez lus, ce n'est peut-être qu'un pour mille de ce qui a été écrit, et ce qui a été écrit c'est encore une toute petite partie de ce qui a été parlé. » Il voulait dire par là que ma base, malgré les 4 000 documents, était trop faible pour arriver à des conclusions sûres. Cela m'a effrayé d'abord et j'ai continué à lire et à dépouiller encore mille ou deux mille documents pour contrôler mes résultats. Le résultat de ce second dépouillement n'a apporté aucune donnée scientifique nouvelle, c'est-à-dire tous les documents n'ont fait que confirmer ce que je savais déjà. Ceci m'a fait voir qu'il y a une relation entre les résultats scientifiques et les données quantitatives. Souvent nous n'avons que peu d'attestations mais chaque attestation augmente la valeur scientifique de ce que je peux en tirer. Mais au fur et à mesure que les attestations se multiplient, la valeur scientifique de chacune diminue. Donc, de l'information à un stade qui du point de vue scientifique mieux vaut restreindre, sélectionner l'information au niveau où elle est scientifiquement utile que de l'augmenter indéfiniment.

M. Quemada : Je suis absolument convaincu que l'information scientifique à partir d'un certain niveau ne s'enrichit en rien des accumulations de données que regroupent les organismes auxquels vous faites allusion. Mon ennui, c'est d'avoir à vous parler des sciences tandis que le genre paralittéraire, auquel appartient un dictionnaire culturel comme *le Trésor de la langue française,* pose des problèmes très différents. Sur le statut linguistique des unités lexicales dans le français du XIX^e et du XX^e siècle, je suis tout à fait d'accord avec vous, il y aurait besoin d'un bagage minime. Mais je pense qu'un dictionniare qui se prétend littéraire, qui prétend avoir dépouillé les grands auteurs des XIX^e et XX^e siècles va mécontenter très sérieusement les spécialistes qui de Balzac, qui de Victor Hugo, qui de Baudelaire, qui de Lamartine, et prenons des contemporains, qui de Camus et de Valéry, ne vont pas retrouver cette fois le passage extraordinaire, la citation merveilleuse de Camus ou de Sartre ou la citation à mot clé, par exemple le mot « existence » chez Sartre. Il y a certainement dix lignes de Sartre où « existence » est en corrélation avec tout un ensemble de termes occurrents par hasard dans ce contexte et qui donnent à la notion « existence », « existentialiste » son poids exceptionnel. Je pense que certains

philosophes seront très fâchés de ne pas trouver cela dans le dictionnaire, de ne pas trouver les témoins, disons de la langue dite littéraire, philosophique, culturelle, etc. C'est là que se situe le vrai problème. Nous ne doutons pas que les fameux choix très aléatoires, faits dans la masse des documents du *Trésor,* fournissent une documentation surabondante pour résoudre le problème purement linguistique, mais non pas pour les problèmes qu'envisageait finalement ce genre très hybride qu'est un dictionnaire de la langue littéraire.

M. Zampolli : Je suis tout à fait d'accord avec le professeur Baldinger, mais je dois donner une petite précision. Quand je parlais d'exhaustivité, je ne parlais pas d'exhaustivité dans le corpus, mais d'exhaustivité dans le lexique et dans la classe linguistique au niveau du lexique. Je sais bien que cela pose des problèmes parce qu'il y a la discussion sur le lexique d'une langue. Est-ce que le lexique est infini ou non, ouvert ou fermé ? Pratiquement, il est fermé ; absolument, il est ouvert. Je ne veux pas prendre position pour une école linguistique ou pour une autre, mais la conception des générativistes est une conception d'exhaustivité, c'est-à-dire c'est génératif, seulement si vous pouvez engendrer tout. Pour engendrer tout vous avez besoin d'un côté, de règles grammaticales et d'un autre côté, d'un lexique qui, en tout cas, peut être en liaison avec ces règles. Il faut alors que la classification soit exhaustive, cela revient à dire que sans lexicologie la grammaire générative n'a pas de sens.

M. Rey : Je crois qu'il y a un point qu'il ne faut pas oublier dans cette discussion, c'est qu'on est constamment au niveau d'un passage entre la linguistique du discours et la linguistique de la langue et qu'il ne faut surtout pas confondre les deux aspects de la question. La statistique, en tant qu'elle parle de fréquence et non pas de probabilité, se situe toujours naturellement dans la perspective du discours et de la parole saussurienne. C'est à ce moment-là que se posent les problèmes évoqués par M. Baldinger sur l'utilisation et l'utilisabilité des éléments dont on dispose. C'est une question d'échantillonnage pensent les statisticiens. Le problème qui a été soulevé par M. Baldinger est en effet une question qui porte sur la valeur de l'échantillonnage. Si on savait à priori qu'elle était la valeur de l'échantillonnage, on pourrait toujours répondre qu'on a une courbe asymptotique, disons à l'horizontale, et quand on arrive au seuil de l'horizontale, on n'a plus besoin d'information. C'est vrai théoriquement, mais en fait on a autant de courbes différentes que le choix des échantillons est plus ou moins pertinent. On ne sait pas à l'avance quelle sera la pertinence du choix des échantillons. Mais si on s'était trouvé dans une situation différente, c'est-à-dire si le chercheur n'avait pas été M. Baldinger, mais disons un de ses jeunes étudiants, il est très vraisemblable qu'avec le même nombre d'occurrences dépouillées, les résultats auraient été beaucoup moins saturés, beaucoup moins complets. Donc, en fait, le problème qui se pose là ne porte pas seulement sur la quantité d'informations accumulées, mais beaucoup plus et surtout sur la qualité de l'échantillonnage par rapport au corpus total envisageable. Je ne sais pas quelles sont les réponses des informations. Est-ce qu'il y a un traitement préalable ou un choix préalable ? C'est le problème qui a été soulevé très souvent pour *le Trésor de la langue française.* On a un très grand nombre d'unités dépouillées, c'est intéressant bien sûr, mais ce qui est encore plus intéressant c'est de savoir où ces unités ont été dépouillées ? Il se peut qu'un seul exemple — et je reviens au point que M. Quemada a illustré tout à l'heure — apporte une information énorme alors qu'un million d'exemples ne va plus apporter d'information du tout. Il est certain qu'on retrouve là l'opposition entre mots lexicaux et mots grammaticaux. Plus un mot est thématique, plus le mot est rare, plus l'information risque d'être importante. Au contraire, quand il s'agit de dépouillement de mots grammaticaux, quand on arrive, pour une zone synchronique donnée, à cent mille ou deux cent mille, d'une part, on n'a pas plus d'information sur la préposition « de » ou sur une conjonction, et d'autre part, le malheureux rédacteur qui va être contraint de rédiger l'article « de », et qui voudrait utiliser tout ce que peut lui donner l'ordinateur du T.L.F., évidemment ne pourra pas les utiliser parce que cela ne sera pas utile. Mais quand il s'agira d'un mot thématique extrêmement rare, le nombre d'occurrences sera naturellement beaucoup plus faible, et chacune des occurrences pourra apporter une information énorme. Ce sont les deux pôles justement de la lexicalité et de la grammaticalité où les problèmes que pose l'automatisation sont complètement différents.

M. Quemada : Je serai tout à fait bref dans la mesure où je suis tout à fait d'accord avec vous. Le corpus du T.L.F. ne permet aucun cas aussi précis, car les règles de sélection sont avec des guillemets « impertinentes ».

M. Zampolli : Je pense que la statistique, la probabilité, c'est quelque chose que l'on pourrait situer peut-être dans la norme. L'on sait que le système du professeur Coseriu est la norme, je ne dirais pas que c'est seulement un fait de discours. Si une construction est plus fréquente, ce n'est pas un hasard. Si nous prenons, par exemple, un échantillonnage d'opinion électorale, on peut très bien connaître les résultats puisqu'il y a des règles bien connues de la statistique. Vous savez que dans cette population, vous avez tant de femmes, tant d'hommes, tant de petits, tant de grands, à peu près tant de pauvres et tant de riches, etc. Pour une langue vous n'avez pas ces données, il faut donc construire un schéma.