LES DÉPOUILLEMENTS ÉLECTRONIQUES: QUELQUES PROBLÈMES DE MÉTHODE ET D'ORGANISATION

par ANTONIO ZAMPOLLI

Si l'initiative de l'Index Thomisticus marque l'introduction de cartes perforées, il est beaucoup plus difficile de dater, sur le plan international, l'introduction des ordinateurs dans les dépouillements de textes. Il est cependant sûr qu'aux Congrès de Tübingen (1960) et de Besançon (1961) on discuta l'opportunité de leur emploi à la place des machines UR. On traça alors un programme de travail avec lequel il est intéressant de confronter, encore aujourd'hui, l'état des choses.

En effet si nous pouvons constater avec satisfaction que dans de nombreux pays la variété des applications, les dimensions des projets et le perfectionnement des méthodologies ont depuis bien longtemps dépassé les prévisions les plus optimistes de ces congrès, il est également vrai que quelques-uns des points les plus importants de ce programme initial doivent encore être réalisés, de sorte que certains problèmes, les problèmes d'organisation par exemple, déjà reconnus comme cruciaux dans les années 60-61, n'ont pas encore été résolus.

Nous avons à ce sujet quelques idées sur des directions de recherche qui pourraient contribuer à les résoudre.

I - *Echange des textes en machine redeable form.* — La phase de préparation des textes en *machine redeable form* représente l'un des plus gros obstacles pour les dépouillements électroniques, à cause du prix de revient, du temps requis par les diver-

ses opérations (perforation — ou ses équivalents —, contrôle, correction) et à cause des problèmes souvent soulevés par la définition des éléments du texte à enregistrer et par leur codification.

En plus des solutions de type technique (collaboration avec les maisons d'édition qui utilisent le type-setting ou la photocomposition; exploration des possibilités offertes par les lecteurs optiques; développement de systèmes *d'editing* appropriés pour la correction, etc...) nous croyons surtout à des solutions au niveau de l'organisation qui puissent assurer l'échange des textes élaborés par des chercheurs ou des équipes différentes.

Sur le plan national la « Divisione Linguistica » du CNUCE a réussi à réaliser presque entièrement cet objectif, en proposant des standards d'enregistrement et d'élaboration des textes qui sont aujourd'hui pratiquement adoptés par tous les projets en cours en Italie. Nous avons mis au point pour le dépouillement des textes un ensemble de procédures et de programmes qui peuvent être appliqués à n'importe quelle langue pourvu qu'il soit possible de la transcrire en alphabet — pour produire les résultats habituels: indices, concordances, fiches lexicales, tableaux statistiques de type divers, etc. On peut affirmer que si un chercheur soumet un texte à la Divisione Linguistica, celui-ci peut être introduit immédiatement dans le cycle élaboratif, car tant le système de codification pour représenter le texte et son analyse que les programmes d'élaboration ont des caractéristiques de généralité suffisantes et adéquates.

Il en résulte que le coût de l'élaboration, le temps et les efforts d'organisation sont beaucoup plus bas qu'ils ne le sont en moyenne dans les autres pays; ceci permet l'accroissement du nombre des textes soumis au dépouillement électronique et pousse les instituts et les chercheurs italiens à s'adresser au CNUCE, passant outre à la tendance au particularisme.

En outre, la présence d'un grand ensemble de textes opérables avec les mêmes programmes et comparables entre eux du point de vue de l'analyse quantitative, grâce à l'uniformité des critères scientifiques et techniques d'élaboration, offre d'importantes possibilités aux études de statistique linguistique pour une révision de ses actuels fondements théoriques. Cette révision concerne aussi, de très près à mon avis, tout le secteur des recherches statistiques sur des textes philosophiques.

On se heurte alors au pressant problème d'éviter la duplication des projets et d'assurer la coopération entre les centres qui opèrent dans les différents pays. Quelques initiatives sont aujourd'hui en cours; elles sont patronnées par divers organismes scientifiques ou politiques et elles se proposent de constituer des banques internationales de données linguistiques. Les projets varient aussi bien pour les solutions techniques et l'organisation proposées, que pour l'ampleur des objectifs qui sont envisagés: on passe, à travers une série d'extensions successives, de l'idée de service international centralisé de secrétariat (qui aurait pour objet de tenir à jour une liste de *machine readable texts* et un carnet d'adresses des personnes intéressées, et de recevoir et de trier les demandes d'exemplaires de textes inclus dans la liste) au projet d'un réseau de centres nationaux qui adopteraient les mêmes standard et les mêmes programmes d'élaboration.

Notre attitude est déterminée par quelques considérations issues d'expériences multiples dans ce secteur.

La simple réunion et diffusion des informations (liste des oeuvres dépouillées, critères généraux de dépouillement, caractéristiques techniques etc...) est en soi très utile. Mais l'expérience apprend que les difficultés d'ordre technique et d'organisation obligent souvent à renoncer à l'utilisation d'un texte déjà disponible dans un autre centre, et l'on doit le reperforer pour soi. Même si l'on a résolu les problèmes de copyright, la conversion d'un système d'enregistrement à un autre n'est pas toujours simple. Dans certains cas le type de support sur lequel

a été enregistré le texte ne peut pas être accepté directement par le système de calcul, à cause de l'incompatibilité des caractéristiques techniques et l'on doit utiliser un dispositif spécial qui serve de médiateur. La description du format, des codes et la définition de la nature des informations enregistrées doivent être des plus méticuleuses et détaillées. En fait la documentation disponible se révèle très souvent insuffisante pour les programmeurs qui doivent écrire le programme de conversion. Enfin la préparation d'un tel programme pour un seul texte ou pour un nombre restreint de textes peut revenir plus cher qu'un travail de reperforation. Par ailleurs, une fois que l'on a établi le programme de conversion, il n'est pas dit que les échanges entre deux équipes qui travaillent dans des centres différents soient assurés une fois pour toutes. Il peut arriver qu'un centre modifie son propre système de calcul et que l'on doive, en conséquence, changer radicalement le système de codification des textes. Il est bien sûr nécessaire, dans ce cas, de refaire le programme de conversion. Les changements de ce type concernent surtout les équipes de lexicographes qui utilisent les ordinateurs des centres de calcul « general purpose »; ces centres ne tiennent en effet guère compte des exigences des usagers humanistes lorsqu'ils décident de développer leur équipement.

Par ailleurs le projet de créer un réseau de centres nationaux qui s'engageraient à adopter un standard unique semble pour le moment utopique. Les problèmes sont souvent indépendants de la bonne volonté de coopérer: les *hardware* peuvent être incompatibles; le système d'enregistrement et d'élaboration peut être étroitement lié à des dispositifs spéciaux, que possède un seul centre; un centre peut déjà disposer d'un ensemble si vaste de textes et d'une série si considérable de procédures et de programmes qu'il serait antiéconomique de les convertir pour adapter les standard internationaux; le codage d'un centre spécialisé dans une langue peut être optimisé par rapport à cette

langue, au point qu'il ne serait guère opportun de les remplacer par un standard plus généralisé, etc...

Nous avons proposé une solution qui semble actuellement la seule concrètement réalisable. Comme elle a été acceptée et choisie par quelqus-uns des organismes internationaux qui patronnent la coopération internationale, nous sommes en train de travailler à sa réalisation.

Il faut avant tout profiter du fait qu'il existe des centres spécialisés dans l'élaboration des données linguistigues qui exécutent, en plus de leurs propres recherches, des projets de dépouillement à la demande de chercheurs de la même région ou d'une nation tout entière. Ces centres ont tout intérêt, pour augmenter la potentialité des services qu'ils offrent, à être en mesure de procurer et d'utiliser pour leurs usagers les textes disponibles dans d'autres centres. Justement parce qu'ils représentent plusieurs équipes qui ont des intérêts communs, ces centres peuvent destiner une partie de leurs ressources aux opérations qui rendent possible l'échange des matériaux: un service de secrétariat qui recueille et distribue les informations, la rédaction des programmes de conversion et leur mise à jour continuelle pour les adapter aux éventuels changements qu'introduisent les centres « partner »; l'acquisition des éventuels dispositifs spéciaux pour assurer la compatibilité des supports d'enregistrement; la mise à jour technologique pour suivre le rapide développement des systèmes de calcul; etc...

La constitution d'un réseau international, en attendant la création des *networks* de computers physiquement reliés les uns aux autres, ne peut être réalisée, à notre avis, qu'à partir d'une série d'accords bilatéraux entre des couples de centres spécialisés, nationaux ou régionaux, qui s'engageraient à créer toutes les conditions (techniques, juridiques et administratives) propres à l'échange des matériaux. Notre travail, dans un proche futur, consistera avant tout à identifier les centres étrangers qui présentent les qualités requises et qui ont la volonté d'échanger avec

nous et avec nos usagers les textes en *machine readable form.* Nous écrirons ensuite les programmes de conversion entre leur codification et la nôtre et inversement, en envoyant, si nécessaire, nos programmeurs apprendre sur place les règles de codification. Nous ne savons pas encore s'il convient ou non de tenter de définir un langage suffisamment puissant pour décrire les divers systèmes de codification, de façon à ne pas devoir écrire un programme pour chaque paire de systèmes. Le fait est qu'entre ces systèmes il n'y a pas seulement les codes qui varient: souvent la nature des informations représentées varie elle aussi. Nous essayerons, en premier lieu, d'établir l'inventaire le plus complet possible.

Les problèmes inhérents au *copyright* ne doivent pas être sous-évalués. Les sociétés constructrices d'ordinateurs les étudient depuis longtemps en ce qui concerne le *software,* tandis que l'étude des problèmes relatifs aux données est moins développée. A ce propos, l'attitude de la « Divisione Linguistica » du CNUCE est la suivante. Si un chercheur demande de pouvoir utiliser un texte élaboré chez nous par un usager, nous demandons avant tout l'autorisation de cet usager, puis nous exigeons la condition que ce dernier soit cité dans les éventuelles publications.

En conclusion nous ne poursuivons pas, pour l'instant du moins, le projet d'un *standard* unique qui devrait être adopté dans tous les pays, mais, en notre qualité de Centre où se font la quasi-totalité des recherches de linguistique computationnelle en Italie, nous essayons d'assurer toutes les conditions nécessaires à l'échange des informations et des matériaux avec un nombre de plus en plus grand de centres d'autres pays. Nous espérons qu'en trois ans de temps nous pourrons réaliser cet accord avec environ quinze des principaux centres européens, américains et canadiens.

Naturellement, lorsque cela nous semble convenable et techniquement possible, nous utilisons les *standard* et les procé-

dures des autres centres et, inversement, nous suggérons l'adoption de nos *standard*. En ce qui concerne le premier cas nous pouvons citer l'exemple de la collaboration entre le CNUCE, le Centro di Studi Egeo-Anatolici et le Department of Eastern Languages de la UCLA de Los Angeles. L'UCLA élabore en ce moment un corpus de textes babyloniens et a mis au point des standards et des procédures très avancées. Le CNUCE a englobé telles quelles dans ses propres systèmes les procédures de l'UCLA qui sont utilisées par le Centro di Studi Egeo-Anatolici pour élaborer un corpus de textes assyriens. On produira, comme résultat final, une banque de données linguistiques assyrobabyloniennes, analysées et élaborées avec une complète uniformité des critères scientifiques et techniques. De cette façon les dépenses et le temps de réalisation de la banque sont pratiquement diminués de moitié.

Pour donner un exemple du second cas nous pouvons citer la banque des données de l'italien. Etant donné que notre banque des textes italiens est de loin la plus riche et la plus vaste, il nous semble juste, chaque fois que nous apprenons qu'un Institut étranger veut entreprendre un projet de dépouillements sur des textes italiens, de nous mettre en contact avec cet institut et de lui suggérer, si la technologie dont il dispose le permet, d'utiliser nos codages et nos procédures. Je dois dire que, jusqu'à présent, nous avons toujours rencontré des instituts parfaitement disposés à collaborer dans cette direction, et cela tout particulièrement en ce qui concerne les centres canadiens et américains.

Nous coopérons aussi à deux autres importantes directions de travail: définition d'un format international pour l'échange (et non pas pour l'élaboration concrète) des textes, des dictionnaires et des données lexicographiques en général; la rédaction d'un 'textbook' avec la définition d'un ensemble de principes et de règles à suivre dans l'encodage des textes en 'machine readable form', que les Associations Internationales devraient recommander aux chercheurs 'débutants' dans le domaine du 'texte processing'.

II - *Dictionnaire de machine et lemmatisation.* — L'exigence de normaliser au moins en quelque sorte les critères de dépouillement se fait sentir même au niveau de l'analyse linguistique du corpus. Cette exigence est chaque jour d'autant plus dramatique que le nombre des textes élaborés électroniquement ne fait qu'augmenter. Ces considérations rejoignent la discussion sur l'opportunité d'insérer, ou non, une phase de lemmatisation ou, plus généralement, l'analyse linguistique dans la procédure de dépouillement. Cette controverse n'est certes pas nouvelle et chacune des deux solutions a ses partisans.

Les arguments contre la lemmatisation sont d'ordre soit pratique soit scientifique. L'ensemble des opérations communément réunis sous le nom de lemmatisation exige une série d'interventions humaines qui interrompent le rythme entièrement automatique des élaborations, augmentant ainsi considérablement le temps et le prix de revient du dépouillement. De nombreux chercheurs préfèrent produire et publier des index et des concordances dans lesquels les exposants ne sont pas des unités définies selon des critères linguistiques, mais sont de simples *formes graphiques*. La rapidité de l'ordinateur est ainsi entièrement exploitiée et l'on peut produire rapidement, à moindres frais, une grande quantité de dépouillements qui, une fois divulgués, peuvent rendre des services indéniables à un vaste groupe de chercheurs. Chaque chercheur exécutera par la suite, pour lui-même, l'analyse de la partie de la documentation linguistique qui l'intéresse, selon les critères qui lui conviennent le plus.

Une certaine analyse et classification des matériaux lexicaux semble toutefois très souvent indispensable avant la conclusion des dépouillements, en particulier lorsqu'il s'agit de dépouillements de grands corpus, par exemple pour la rédaction de vastes dictionnaires historiques. Cela pour éviter que les usagers du dépouillement, dans notre exemple les rédacteurs du dictionnaire, soient submergés par la quantité des données à choisir et à ordonner. On se demande inévitablement 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.

Une des tâches de la linguistique computationnelle est donc, semble-t-il, de fournir les instruments qui rendent plus simples, plus économiques et plus sûres les opérations d'analyse.

Le dictionnaire de machine ou lexique automatique (L A) est le plus simple de ces instruments. Comme on le sait, les premiers L A furent mis au point pour la traduction automatique du russe à l'anglais et vice versa, dès 1946; au début il y eut même des personnes qui estimèrent qu'un bon L A était une condition nécessaire et suffisante pour traduire automatiquement. Les premiers textes de linguistique computationnelle dédient une grande place aux systèmes de compilation et de consultation d'un L A. Toutes ces études ont produit des systèmes très connus et raffinés pour la gestion et la consultation d'un L A.

Toutefois on ne doit pas s'étonner si, malgré cela, je place l'utilisation des L A parmi les développements possibles de la lexicographie assistée par les ordinateurs. En réalité il y a très peu d'entreprises lexicographiques et en général très peu d'auteurs de dépouillements lexicaux et statistiques qui aient adapté un L A pour rendre automatique, ou tout du moins semi-automatique, la lemmatisation.

Nombreux sont ceux qui objectent que cette lemmatisation réalisée au moyen d'un L A selon la procédure habituelle, fait courir le risque de commettre de graves erreurs. La plus grave de ces erreurs serait qu'une forme effectivement homographe dans le système linguistique auquel appartient le texte à lemmatiser ait été insérée, au contraire, en tant qu'univoque dans le L A, soit par erreur matérielle, soit par insuffisance de connaissance de la langue. Dans ce cas l'ordinateur lemmatiserait directement en tant qu'univoques les occurences de cette forme dans le texte, sans les soumettre à l'examen du linguiste. Plus

la couche de langue étudiée s'étend diachroniquement, plus elle régresse dans le temps vers des systèmes qui ne sont pas entièrement présents à la compétence de celui qui compile le L A, et plus ce risque est grave.

L'obstacle est toutefois surmontable si, au lieu de considérer comme définitive la lemmatisation des formes univoques selon le L A, on en imprime les contextes lemmatisés, afin que le linguiste puisse contrôler les lemmes attribués par le L A. Des expériences appropriées ont démontré que dans ce cas on économiserait toujours un temps considérable (70% environ) par rapport à la procédure de lemmatisation manuelle.

Etant donné que le grand nombre de transcriptions nécessaires dans la procédure artisanale de lemmatisation est évité par l'emploi du L A, on obtient une réduction considérable du risque d'erreurs casuelles. Le L A fait en même temps fonction de système de référence, assurant ainsi une certaine cohérence de comportement dans les cas où la présence d'alternatives systématiques exige que la formulation du lemme se fasse selon des règles fixées. Le dictionnaire de machine fonctionne en tant qu'enregistrement des décisions prises, et comme il peut être imprimé rapidement ou consulté à travers un écran, il fournit à chaque instant un tableau complet des formes examinées et des traitements adoptés. A côté de cette fonction normalisatrice, qui est d'autant plus importante que le corpus soumis au dépouillement est plus étendu et que les lemmatisateurs sont plus nombreux et davantage distribués dans le temps, le L A exerce également un ensemble de fonctions collatérales; C.A. Mastrelli a par exemple suggéré, pour l'ACCADEMIA DELLA CRUSCA, de dater automatiquement les formes dans le corpus choisi.

Les objections d'ordre scientifique sont formulées de diverses façons, mais elles peuvent être résumées, d'une manière simplificatrice, par l'affirmation que la lemmatisation substitue à l'objectivité du critère d'identité des données au niveau graphémique, la subjectivité de l'interprétation personnelle, qui se

produit nécessairement dans le cadre actuel de la pluralité des théories linguistiques changeantes et en général non entièrement spécifiées.

Analyser un texte signifie avant tout y reconnaître les unités du système linguistique et leurs rapports syntagmatiques. L'inventaire et la définition des unités ainsi que la reconnaissance de leurs rapports dépendent d'une théorie linguistique. Les théories linguistiques, surtout aujourd'hui, semblent changer assez rapidement. Il est inévitable que celui qui dépouille des textes ou un corpus de grande dimension, se demande si les résultats de son travail seront liés à une théorie spécifique au point d'être mal utilisables ou même non interprétables dans une théorie différente. Les théories actuelles se trouvent le plus souvent développées au niveau de modèles que l'on essaie d'adapter à des sous-ensembles très restreints de langue ou même à l'explication de faits spécifiques, et qui ne semblent pas appropriés à la description de sous-ensembles de langue aussi étendus que ceux qui se trouvent dans un texte. Certains ajoutent aussi que l'analyse est nuisible et vicieuse si elle n'est pas poussée au niveau maximal de détail prévu par la théorie. Mais, en ce cas, elle deviendrait pratiquement inexécutable sur un texte de dimensions normales.

Il n'est pas facile de répondre à cet interrogatif qui se pose de façon d'autant plus dramatique que le corpus est plus vaste et donc que le dépouillement est proportionnellement plus coûteux. On a pu le constater au cours des vives discussions qui se sont tenues à ce sujet au *Colloque sur l'indexation maximale* de Strasbourg (1973). Dans tous les cas cette réponse ne peut pas se réduire à un oui ou à un non, mais elle doit au contraire prendre la forme de procédures qui assurent, dans la limite du possible, le dynamisme de l'analyse linguistique du corpus.

Entre l'attitude extrême de ceux qui soutiennent que l'analyse d'un corpus dépend entièrement d'une théorie et n'a de sens que si elle est poussée au niveau maximal de détail prévu par la

théorie, et l'autre cas extrême de ceux qui, à partir de ces mêmes considérations, concluent que l'indexation doit s'arrêter au relevé automatique des unités purement graphiques, en laissant toute analyse à l'usager des indices, se trouvent, à notre avis, des procédés intermédiaires. Ces processus semblent s'appuyer sur une conviction qui, souvent, n'est pas explicitement affirmée, selon laquelle il serait possible de déterminer des unités qui tout en étant définies par des critères linguistiques, sont, pou ainsi dire, neutres par rapport à la plupart des théories lir guistiques. Ces unités pourraient être indiquées et délimitée univoquement même si leur statut, leur classification et leur relations réciproques peuvent varier d'une théorie à l'autre.

Dans ce cas, le L A serait conçu comme un inventaire de ces unités de base, et l'analyse consisterait à les reconnaître dans le corpus. Autrement dit, l'indexation consisterait à relier avec une unité du L A toutes ses occurrences dans le corpus. Nombre de déplacements induits dans l'état, dans les définitions et dans les relations réciproques des unités d'une théorie pourraient être enregistrés dans le L A sans altérer l'inventaire de ses unités, mais en modifiant simplement leurs descriptions. Le corpus se trouverait automatiquement réanalysé selon la nouvelle classification sans qu'il soit nécessaire d'intervenir sur lui.

La validité de cette solution peut être discutée selon de nombreux points de vue. Le point crucial consiste naturellement dans la démonstration qu'il existe des unités de base qui peuvent être traitées par une théorie ou par une autre rien qu'en changeant leur description. Je ne peux pas m'étendre à présent sur cet argument pour le discuter. Je peux toutefois affirmer que notre expérience de dépouillements nous donne la confirmation qu'il est possible de construire, à un niveau très superficiel d'analyse, un inventaire satisfaisant de ces unités ou, du moins, un inventaire où la marge d'erreurs dues à l'adaptation de divers systèmes théoriques serait compensée par la possibilité

d'étendre l'analyse à des corpus de dimensions autrement prohibitives.

La *Divisione Linguistica* du CNUCE a pour projet un « Dizionario Italiano di Macchina » (L A I), dont le noyau comprend un ensemble de 120.000 lemmes environ, obtenu en unissant les lemmes des principaux dictionnaires italiens; ce noyau pourra être ensuite enrichi, surtout par des termes techniques, avec les résultats des dépouillements effectués auprès du CNUCE. Pour chaque lemme nous avons enregistré dans une première phase d'élaboration l'appartenance éventuelle à des secteurs particuliers du lexique (dialectaux, techniques, archaïques, etc.), 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, également un dictionnaire de formes.

Dans une seconde phase, nous enrichirons la série des informations en ajoutant à chaque lemme des classificateurs syntaxiques, sémantiques et statistiques.

La structure du L A I et l'algorithme de consultation ont été organisés de façon à concilier deux exigences apparemment opposées: permettre à l'usager de choisir parmi divers types de lemmatisations et en même temps assurer un degré suffisant de comparabilité entre leurs résultats.

En effet une des objections les plus communes contre le L A est que l'on devrait créer un L A différent pour chaque dépouillement, car l'auteur du dépouillement peut avoir envie d'adopter des critères différents dans la lemmatisation selon ses buts et la nature du texte. Nos expériences montrent que les différences de comportement dans la façon de lemmatiser peuvent être ramenées à deux types principaux. Sous une pre-

mière rubrique, on peut regrouper toutes les formes dans lesqelles 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. L'utilisation du L A garantit que chacun de ces cas reçoit le même traitement de lemmatisateurs différents. Par contre, les formes de la deuxième rubrique 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,* ecc.) L'usager pourra adapter le L A I à ses propres exigences, en communicant au programme, au moyen de 'control-cards', l'alternative préférée pour chaque classe. L'important est que, grâce au L A I, toutes les formes d'une même classe soient énumérables, qu'elles soient traitées suivant le même critère, et que le critère soit formulé de manière explicite.

Dans ce cas il est possible de rendre homogènes les résultats des dépouillements. Voyons par exemple ce qui arrive pour les situations qui comprennent la majeure partie des cas.

Soit x une forme graphique, appartenant à une classe Y qui admet les alternatives suivantes:

a) la forme est *homographe* et doit être analysée pour distinguer

1) 3 formes: (x_1) (x_2) (x_3)
2) 2 formes: $(x_4 = x_1 + x_2)$ (x_3)
3) 2 formes: (x_1) $(x_5 = x_2 + x_3)$

b) la forme est *univoque*:

4) 1 forme: $(x_6 = x_1 + x_2 + x_3)$

où par exemple $x_4 = x_1 + x_2$ signifie que dans la seconde alternative à la forme x_4 sont attribuées toutes les occurrences

qui dans la première alternative sont attribuées à x_1 ou à x_2 et seulement celles-ci. Supposons qu'un usager choisisse de lemmatiser un texte *A* selon la deuxième alternative. Il peut la communiquer au programme de consultation, par exemple, avec une fiche de contrôle sur laquelle il perfore le sigle de la classe (*Y*) et le numéro de l'alternative choisie (2). L'ordinateur considérera comme des homographes possibles toutes les formes de *Y* et proposera à l'usager de distinguer les occurrences de chacune d'elles entre x_4 et x_2, garantissant ainsi le fait que toutes les formes de Y sont analysées de manière cohérente dans le corpus.

Si par la suite on veut confronter les résultats du dépouillement du texte *A* avec ceux qui ont été obtenus sur un texte *B* lemmatisé en choisissant pour *Y* l'alternative I, il est évident qu'avant de comparer *A* avec *B* il suffira d'additioner entre elles, pour *B*, les occurrences de x_1 et de x_2. En général, deux textes lemmatisés avec le L A I peuvent être rendus comparables, automatiquement, avec un programme qui, chaque fois qu'il rencontre des formes d'une classe qui ont été distinctes dans un texte et dans un autre non, veille à les unifier dans le premier.

J'ai déjà énuméré autrefois [1] les recherches linguistiques qui, en supposant l'existence d'un inventaire des unités lexicales du système linguistique italien, trouvent dans le LAI un instrument efficace. Je voudrais simplement ici ajouter une considération: un L A dont les mots sont classés de façon adéquate, est l'instrument typique pour effectuer des recherches sur des textes philosophiques qui se proposent des objectifs analogues à ceux de la soi-disant *content-analysis*. A ce sujet je renvois à la communication de M. Delatte. Le projet que nous avons à l'étude

[1] A. ZAMPOLLI, *La section linguistique du CNUCE*, in A. ZAMPOLLI (ed.) *Linguistica Matematica e Calcolatori*, Firenze, 1973, pp. 133-199, et in A. ZAMPOLLI, *Humanities Computing in Italy*, in « Computers and the Humanities », VII (1973) 6, pp. 343-360.

porte pour le moment sur l'analyse du contenu des articles politiques des journaux.

III - *Distinction automatique des homographes.* — Pour compléter l'automation de la lemmatisation il faudrait pouvoir distinguer automatiquement les homographes signalés par le L A. Ce problème a été et est affronté depuis longtemps dans les projets de traduction automatique, tandis que je connais seulement de rares tentatives de la part des entreprises lexicographiques. A Nancy l'équipe du Trésor avait commencé à étudier des algorithmes pour les homographes de très haute fréquence et nous sommes en train de faire la même chose à Pise. Des tentatives analogues sont en cours à l'Académie Royale Espagnole.

En ce qui concerne l'homographie fonctionnelle (du type substantif-verbe: *faccia*) on se propose un *parser* syntaxique.

Le principe est évident. Le L A associe à un homographe de ce type plusieurs analyses grammaticales distinctes, une pour chaque fonction syntaxique différente que l'homographe pourrait exercer dans l'énoncé. Si le *parser* réussit et donne 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.

Pour l'homographie de type radical, du type *mozzo* (*essieu de la roue* et *mousse sur un navire*), qui ne comporte pas d'analyses grammaticales différentes, on proposait jusqu'en 1966-67 un *parser* à niveau sémantique componentiel. On pensait par exemple classer tous les mots présents dans le L A selon des catégories ou tout au moins des composantes sémantiques et formuler des règles qui spécifieraient la possibilité ou l'impossibilité de relation, de sélection, de cooccurrence entre les diverses catégories sémantiques, au moins dans certaines positions de la structure syntaxique. Cela, dans le but plus ou moins explicite,

le construire un modèle sémantique global, un réseau de rapports reliant, à la limite, tous les mots du lexique.

Aujourd'hui la situation a changé et l'on a tendance à ne pas séparer les deux moments du *parser*. J'ai décrit ailleurs[2] l'état des procédures pour l'analyse syntaxique et sémantique. Il suffira d'observer ici que si les systèmes les plus récents (Woods, Winograd, etc.), réussissent à traiter une grande variété de structures syntaxiques complexes, il sont en général limités à un sous-ensemble lexical très restreint et leur fonctionnement suppose que les énoncés qui sont soumis à analyse se réfèrent à un sujet précis, connu à priori, à une partie bien précise de la réalité, dont la connaissance est donnée au système comme une représentation formelle explicite.

Les occasions de rencontre entre lexicographes et lexicologues, absorbés par le dépouillement de vastes corpus de textes, et les linguistes computationnels absorbés par l'étude des soi-disant systèmes intégrés d'analyse linguistique, ont manqué même dans le passé le plus récent. On a cherché délibérément ces derniers temps à créer des occasions de rencontre et de discussion: en effet les chercheurs qui, comme le soussigné, ont des expériences dans les deux domaines, sont convaincus de la possibilité et de la nécessité d'échanges méthodologiques.

Au cours de la session 1972 de l'Ecole Internationale d'Eté de Pise *Mathematical and Computational Linguistics* et de la *International Conference on Compulation Linguistics* (Pise, 1973) les auteurs des systèmes de *parser* les plus récents qui ont été interrogés par les lexicographes et les lexicologues sur la possibilité d'étendre leurs modèles jusqu'à inclure des sous-ensembles de plus en plus vastes, et à la limite toute la langue d'usage commun, ont exprimé des jugements opposés sur la possibilité théorique d'une telle

[2] A. Zampolli, *Problemi di Linguistica applicata computazionale*, Pisa, 1974.

extension, mais ils ont tous été d'accord sur l'impossibilité pratique de la réaliser concrètement, au moins pas avant plusieurs dizaines d'années. Ceci parce que, à la différence des systèmes de *question-answering*, de *speechunderstanding*, etc..., qui s'occupent de sous-ensembles de langue naturelle extrêmement réduits et d'une certaine façon le plus souvent déjà formalisés, les lexicographes soumettent au dépouillement des corpus de textes largement distribués aussi bien diachroniquement que synchroniquement (sujet, registre d'emploi, genre littéraire, etc...). On doit donc conclure que la solution automatique de toutes les homographies d'un texte est, aujourd'hui tout au moins, un but utopique. Mais on ne doit pas exclure, ni renoncer à automatiser *en partie* la solution de l'homographie. Même s'il est rare que soit étudiée ou tout du moins que soit divulguée l'efficience des systèmes existants de *parser,* nous possédons néanmoins quelques données indicatives selon lesquelles sur des textes de langue anglaise, choisis sans restriction particulière, il serait possible de résoudre automatiquement, en incorporant quelques-unes des plus récentes découvertes de systèmes intégrés, 60% des homoghaphies. Avec des algorithmes *ad hoc,* qui, tout en incorporant ces découvertes, ne visent pas à reconnaître la structure complète de la phrase, mais se proposent de résoudre surtout l'homographie fonctionnelle qui, comme on le sait, provient en grande partie d'un petit nombre de mots à très haute fréquence, il semblerait possible d'atteindre et dans certaines langues de dépasser 80%. Résultat très appréciable car, à la différence de systèmes intégrés où est important le processus d'analyse en lui-même, dans les dépouillements il importe essentiellement de réduire le travail humain.

Probablement les premiers pas concrets devraient être faits en direction d'une interaction homme-machine pour la résolution de l'homographie, dans l'attente d'atteindre un état de l'art où le programme devrait exécuter le parsing des phrases pour les-

quelles il est adéquat et devrait au contraire exiger la collaboration du linguiste pour les phrases qui dépassent ses capacités.

Sur l'écran apparaît la forme à analyser. A côté d'elle apparaissent les diverses propositions d'analyses fournies par le L A et par le *parser* avec des numéros progressifs. 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 occurrences, le lemmatisateur contrôle l'exactitude de l'analyse choisie. Autrement il l'effectue lui-même en choisissant, au moyen de la touche ou du *light-pen,* parmi les analyses proposées par le L A. Au cas où aucune des analyses proposées ne conviendrait à une certaine occurrence, le lemmatisateur ajoutera la nouvelle analyse à celles déjà existantes dans le L A.

Au CNUCE nous sommes en train d'étudier un projet pour la construction d'un *parser* syntaxique de l'italien qui ne se propose pas de produire, pour l'instant du moins, des analyses complexes au niveau de la structure profonde des énoncés. Ce *parser* vise plutôt à produire une analyse syntaxique superficielle, qui ne veut pas nécessairement reconstruire l'indicateur syntagmatique de la phrase entière, mais qui peut éventuellement se limiter à reconnaître les syntagmes du type groupe nominal, groupe verbal, etc., sans prétendre les réunir en une structure unique. Nous sommes convaincus que ce type d'analyse permettra de distinguer correctement la plupart des homographes fonctionnels comme le montrent les travaux de Ross, Milio et Vauquois.

IV - *Selection des matériaux lexicaux.* — Le thème de la sélection domine le processus de compilation des lexiques, des concordances et surtout des dictionnaires, en particulier des dictionnaires historiques.

Au départ les compilateurs doivent choisir un corpus de textes qui puisse être considéré comme représentatif du corpus de textes théoriquement disponibles dans la couche de langue sur laquelle porte la recherche. On ne peut nier l'opportunité

d'une sorte de *feedback*. Au fur et à mesure que leur dépouillement fournit, tout en progressant, de nouvelles informations sur la structure de l'échantillon examiné, il devrait être possible de modifier le corpus lui-même, en réduisant par exemple le nombre des textes relatifs à un sous-ensemble de langue ou à des classes de faits linguistiques suffisamment documentés; ou bien, inversement, en introduisant des textes qui sont supposés contenir des phénomènes qui ne soient pas encore parus, etc. Ce type de *feedback,* très peu réalisé jusqu'à aujourd'hui, exige une organisation particulière du dépouillement, pour pouvoir disposer rapidement d'archives constamment mises à jour et rapidement accessibles dans toutes ses parties. Les caractéristiques de telles archives sont les mêmes que celles de la soi-disant *banque des mots*, dont nous parlerons au paragraphe suivant.

Le second processus de sélection consiste à choisir dans le corpus les exemples qui doivent faire partie des archives.

Les lexicographes sont d'accord, en général, avec le calcul de I. Aitken, selon lequel un expert lexicographe peut examiner, en rédigeant des articles de dictionnaire, presque 10.000 fiches-contexte chaque année. A cette vitesse, l'élaboration composée par 10.000.000 de citations demanderait 100 collaborateurs pendant 10 ans. Les archives des grands dictionnaires historiques de ces 150 dernières années, basés sur des dépouillements manuels, comprenaient pratiquement d'un demi-million à 10 millions d'occurrences. Malgré les devis prévus pour le temps, ils les ont presque toujours dépassés de plusieurs dizaines d'années. L'utilisation d'un ordinateur qui permette de recueilir, en un temps relativement court, un nombre d'exemples bien supérieur, pourrait à la limite rendre plus dramatique le problème de choisir les occurrences particulièrement « intéressantes » à insérer dans les archives.

L'Accademia della Crusca et le CNUCE ont mis au point une procédure pour la composition d'archives lexicales qui est encore aujord'hui, d'un point de vue global, parmi les plus éco-

nomiques et les plus pratiques. Le dépouillement électronique prévoit l'enregistrement intégral du texte en *machine readable form* et la production des concordances de toutes les formes du texte. Une équipe de lemmatiseurs lit les concordances par forme en les lemmatisant et en marquant en même temps les occurences retenues dignes d'entrer aux archives. Quelques expériences pour automatiser au moins en partie cette sélection ont été accomplies au CNUCE et, si j'ai bien compris, au T.L.F. de Nancy. Ces expériences sont surtout basées sur des méthodes statistiques et les résultats obtenus sont encore très discutables. Au cours de l'Ecole d'Eté de Pise de 1[illegible] on a amplement parlé des moyens capables d'abréger ces opérations de choix et de les baser, si possible, sur des critères qui ne dépendent pas exclusivement du jugement individuel du lemmatiseur. L'idée de base prévoit l'interaction du lexicographe avec le programme, au moyen d'un écran. L'ordinateur pourrait activement aider le lexicographe: dans le processus de lemmatisation, par la consultation d'un dictionnaire de machine; dans le processus de sélection, par la reconnaissance automatique dans le texte de quelques structures définies auparavant en tant que facteurs importants pour l'acceptation ou l'exclusion des exemples, et par le contrôle statistique continuel des éléments choisis.

Le troisième procédé de sélection appartient à la phase de rédaction proprement dite du dictionnaire. Les rédacteurs doivent analyser les matériaux recueillis dans les archives, choisir les exemples les plus représentatifs et les organiser d'une manière opportune, de façon à illustrer chaque acception, sa distribution et son évolution dans le temps, dans l'espace etc..., écrire la définition et enfin suivre jusqu'à son impression le résultat de toutes les opérations. On peut imaginer un processus de rédaction à l'aide de l'ordinateur.

Le L A fournirait avant tout un premier schéma de la structure de l'article, schéma éventuellement modifiable suivant les nouvelles évidences qui émergeraient de l'examen des archives.

Le rédacteur pourrait faire apparaître par séquence sur l'écran les exemples et au moyen de simples commandes sur le clavier ou avec le *light-pen* il pourrait effectuer en une même phase les opérations suivantes:

— indiquer avec des codes le regroupement des exemples en classes;

— demander à l'ordinateur de présenter à nouveau les exemples groupés selon ces classes;

— indiquer les exemples qui doivent confluer dans le dictionnaire en taillant éventuellement de la façon la plus opportune les exemples qui seraient trop longs;

— insérer en tête de chaque groupe la définition, ou plus généralement les classificateurs pertinents;

— envoyer les matériaux ainsi élaborés (définitions et exemples) à un processus automatique d'impression, par exemple la photocomposition.

Un schéma de ce type a été discuté à l'Ecole d'Eté 1972 à Pise. La *Divisione Linguistica* du CNUCE est en train de développer, en collaboration avec B. Quemada, une procédure interactive qui complète la première expérience présentée en 1972 à l'Ecole d'Eté.

On peut objecter à juste titre à ce schéma qu'il ne serait pas possible d'effectuer, sur l'écran, l'examen synoptique des divers exemples, chose que le lexicographe est habitué à faire en « éparpillant » les fiches-contexte sur son bureau. Certains ont proposé de remédier à cela en mettant deux écrans en parallèle; sur l'un d'eux apparaîtrait le contexte étudié, tandis que sur l'autre le chercheur ferait défiler, à son aise, les autres contextes de la même forme. D'autres ont proposé de recourir à des techniques particulières qui combinent un écran terminal et les microfiches. Il est sûr et certain qu'il faudra encore un sérieux et peut-être un long travail d'expérimentation en étroite collaboration entre les linguistes computationnels et les lexico-

graphes. De toute façon c'est certainement là l'une des plus intéressantes perspectives de développement qui s'insère naturellement dans le projet plus général d'une *banque de mots.*

1.5. *La banque de mots.*

A l'Ecole d'Eté de 1972 se sont dessinées deux interprétations différentes du terme *banque de mots* ou archives lexicales.

D'une part quelques-uns [co]mme J. Bahr l'ont définie comme une sorte de L A où cha[que] mot est accompagné par toutes les informations phonologiques, morphologiques, syntaxiques, sémantiques connues. Quemada et moi-même l'entendions comme de véritables archives de mots extraits de corpus de textes analysés, élaborés, et non imprimés, mais conservés par l'ordinateur et mis à la disposition des linguistes et des lexicographes au moyen des techniques modernes de conversation et d'interaction homme-machine.

Je pense qu'une *banque de mots* doit contenir ces deux éléments à la fois, le L A et les textes, et que le L A aussi bien que le corpus doivent avoir des caractéristiques dynamiques.

Le L A peut constituer l'enregistrement où se déposent les connaissances acquises par des linguistes et des lexicographes sur le lexique et sur ses relations fonctionnelles avec les autres composantes du système linguistique. Le L A fournit ainsi un inventaire des unités du système lexical et de leurs propriétés, lesquelles font office de classificateurs par rapport aux unités du texte. L'unité linguistiquement définie et classée du L A est la clef à travers laquelle on accède normalement aux unités correspondantes du corpus.

Naturellement le L A doit être perméable à toutes les modifications exigées par les nouvelles théories et peut ainsi servir à des comparaisons entre leur efficacité descriptive.

Mais le L A doit pouvoir être intégré ou modifié même selon des données provenant de l'analyse du corpus.

Une des informations qui viennent du corpus au L A est celle de la fréquence. Cela servira à combler progressivement la lacune à juste titre déplorée par J. Rey-Debove. Pour chaque mot, pour chaque acception de ce mot, pour les constructions grammaticales auxquelles elle participe, etc..., la banque des mots devrait permettre l'attribution d'une fréquence d'usage, distincte dans les différents sous-ensembles de textes identifiables sur l'axe diachronique et synchronique.

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. Et c'est là une tâche d'autant plus ingrate que l'analyse est plus détaillée et le corpus plus vaste.

Ainsi se pose à nouveau le problème de réduire la disproportion entre les données produites par les dépouillements et les possibilités élaboratives de l'homme. Nous pouvons seulement énumérer quelques éléments préparatoires pour donner une réponse provisoire.

Il est nécessaire en premier lieu de coordonner les efforts de tous ceux qui travaillent sur une même langue et de créer des banques de données et de programmes à un niveau national ou international. Le centre national qui gère la banque doit être doté de *hardware* et de *software* spécifiquement désignés pour les exigences linguistiques, mais il doit également disposer de systèmes d'imprimerie qui permettent de réaliser des éditions satisfaisantes.

L'analyse des données doit être décentralisée, c'est-à-dire qu'elle doit être accomplie par de petites équipes de chercheurs ou même par des chercheurs isolés, dotés d'équipements modestes, mais qui puissent accéder à la banque d'où ils extraient les textes à analyser et où ils les « redéposent » une fois analysés. L'emploi d'un L A et un ensemble de critères d'analyses formulées explicitement assureraient un niveau acceptable de cohérence entre les divers exécutants des analyses. En effet, la présence

d'un L A fonctionne comme l'enregistrement de toutes les solutions déjà trouvées, et suggère un schéma commun d'analyse. La banque dit quels usages, quelles constructions ont été trouvés, avec quelle fréquence. Elle met en évidence les lacunes des corpus et suggère quand il faut recourir aux compétences des informateurs.

Ce dialogue à trois, L A — corpus — chercheur, semble être le seul système qui permette de passer d'archives statiques, imprimées dans une marée de fiches ou de concordances, à une banque moderne de données lexicales qui fasse fructifier les ressources offertes par la technologie moderne. De cette banque seraient imprimées (de préférence automatiquement avec des systèmes de microfiches reliés à l'ordinateur) seulement quelques parties à la demande des chercheurs. Dans la plupart des cas, elle serait consultée au moyen de terminaux. Pour les raisons que l'on a énoncées il est important d'avoir une banque de mots centralisée pour chaque langue.

On ne peut qu'espérer que la coopération réalisée en Italie ne sera pas gâtée par la tendance au particularisme et au chauvinisme, tendance que dans d'autres secteurs la presse nationale signale justement ces temps-ci. Elle témoigne que l'illusion du pouvoir, donnée par la possibilité de disposer d'un centre de calcul autonome, produit, surtout dans l'administration publique, la duplication des efforts et le morcellement des ressources. De là dérive inévitablement l'incapacité d'approfondir les problèmes et de faire progresser les méthodologies.