ANTONIO ZAMPOLLI

LA SECTION LINGUISTIQUE DU CNUCE

1. LA SECTION LINGUISTIQUE DU CNUCE

En 1966, dans le cadre des travaux de l'Index Thomisticus et du dictionnaire de l'Accademia della Crusca, on commença à utiliser les élaborateurs IBM du Centre National Universitaire de Calcul Électronique de Pise (CNUCE)[1] pour réaliser les dépouillements lexicaux automatiques prévus par le programme de ces recherches.[2] D'autres Instituts suivirent bientôt cet exemple, et en 1968 les projets d'emploi de l'ordinateur dans plusieurs secteurs des disciplines humanistes étaient si nombreux que la Direction du CNUCE décida de constituer une section

---

[1] En mai 1964 IBM–Italie mettait à la disposition de l'université italienne un ensemble électronique du type IBM 7090, attribué sur décision du Ministre de l'Education Nationale à l'Université de Pise.

Ainsi naissait, par une convention signée entre l'Université même et IBM Italie, le *Centre Universitaire de Calcul Electronique* dont le siège se trouverait à l'« Università degli Studi » de Pise.

Toutes les Universités italiennes et étrangères peuvent participer aux initiatives du Centre.

La convention a été renouvelée cette année et a prévu un ordinateur IBM 360/67.

[2] En Italie le premier projet de lexicographie automatisée a été celui de l'Index Thomisticus, commencé par R. BUSA S. J., à Gallarate, en 1949. (R. BUSA, A. ZAMPOLLI, 1968). Pendant l'année universitaire 1959-1960 fut discutée à l'Université de Padoue la thèse de A. ZAMPOLLI, (1960), sous la direction de C. Tagliavini, qui constitue la première tentative d'utilisation des ordinateurs pour le dépouillement et l'analyse d'un texte italien d'un point de vue phonématique, lexical et morphologique. (Cf. C. TAGLIAVINI, 1968). En 1964 furent réalisées sous la direction de A. DURO les premières expérimentations de dépouillement pour le grand *Dictionnaire Historique de la Langue Italienne* dont la réalisation fut confiée à l'Académie de la Crusca par le C. N. R. Le procédé électronique fut conçu par A. Zampolli. Ce procédé constitue la base des programmes pour les dépouillements lexicaux aujourd'hui en usage auprès du CNUCE. (Cf. A. DURO, A. ZAMPOLLI, 1968). En 1965, à l'occasion de l'inauguration du CNUCE à Pise, fut offerte au chef de l'Etat une copie des concordances et des index lexicaux de la Divine Comédie, élaborés électroniquement par C. Tagliavini sur l'initiative d'IBM Italie.

linguistique,[1] qui, d'une part, encouragerait les recherches et les études originales, et qui, d'autre part, mettrait à la disposition des humanistes intéressés, non seulement les élaborateurs électroniques et le personnel technique nécessaire pour l'exécution des opérations, mais, permettrait aussi la consultation scientifique, l'analyse et la rédaction de nouveaux programmes, lorsque les programmes généralisés (utility programs) ne seraient pas suffisants.

Les Instituts Universitaires et de Recherche, italiens et étrangers, qui mènent des recherches dans le secteur humaniste auprès du CNUCE sont aujourd'hui plus de 50 et ils ont enregistré et élaboré électroniquement non moins de 50 millions de mots dans plus de 20 langues [2]. Les avantages scientifiques et économiques offerts par le rassemblement de projets sont remarquables: pensez, par exemple, que l'adoption d'un même système général d'enregistrement et d'élaboration permet d'utiliser des programmes déjà expérimentés tout en laissant à chaque chercheur la possibilité d'utiliser chaque nouveau programme écrit pour un projet et tout en autorisant l'échange entre les divers Instituts. En pratique, un chercheur peut mener ses recherches dans une grande bibliothèque électronique à fonctionnement automatique, plutôt que sur le seul texte enregistré par lui-même. D'autre part, la présence d'un corpus si étendu incite les chercheurs qui entreprennent de nouveaux dépouillements, à adopter les « standards » que nous proposons. Ils sont le résultat d'expériences collectives, menées depuis plus de dix ans, et comportent un schéma d'enregistrement et d'élaboration bien formulé. Ceci exige que l'on ne néglige pas des informations qui, bien que n'étant pas immédiatement nécessaires pour la recherche spécifique proposée, sont toutefois indispensables pour garantir l'utilisation des matériaux de la part des autres chercheurs.

---

[1] La structure de la Section Linguistique est la suivante:

Le Directeur de la section a sous ses ordres d'une part 9 chercheurs diplômés en Sciences Humaines et d'autre part un responsable des opérations qui dirige 6 programmeurs, 2 perforatrices à plein temps et un groupe de 12 perforatrices à mi-temps.

Actuellement, aux élaborations linguistiques sont consacrées 8 heures de travail sur IBM 360/30 64K (doté de 4 unités bande, de 2 unités disque, d'un lecteur-perforateur, d'un transcripteur rapide 1403/n. 1 et d'un trascripteur-lecteur de fiches 1404) et 16 heures de travail sur IBM 1401/16K. En outre on peut avoir accès, en en faisant la demande, à l'IBM 7090-100 K doté de 12 unités bande et utilisé seulement pour classer et calculer les formules statistiques les plus complexes. Il est prévu pour l'été 1971 la mise en service d'un système IBM 360/67/1024 K qui entraînera certainement un changement au niveau des modalités d'opération.

[2] Le rapport périodique du CNUCE contient chaque année le compte-rendu des travaux en cours et de leurs progrès.

Il sera important de maintenir cette unité scientifique et technique même lorsque, grâce à l'installation d'un nouveau ordinateur IBM 360/67, il sera possible et conseillé aux quelques Instituts intéressés, de faire au moins une partie des élaborations, et particulièrement celles qui exigent un dialogue entre chercheur et ordinateur, au moyen de leurs propres terminaux reliés au 360/67.[1]

---

[1] Excepté le problème de la lemmatisation, que certains évitent en élaborant des index et des concordances sous forme graphique et non sous forme de lemmes (Cf. par exemple M. ALINEI, 1968 et l'article de F. DE TOLLENAERE dans ce volume), il existe d'autres problèmes d'ordre technique et pratique à propos desquels M. KAY (1967, p. 170) écrit: « For every person who has successfully completed a linguistic or literary project using a computer, there are probably three or four who (...) have embarked on such projects and who, after a few frustating weeks, months, or even years, have abandoned them in disgust (...). What nobody tells the aspiring computer user is that these machines can all too easily replace one kind of drudgery by another ». Les problèmes les plus fréquents semblent être: obtenir les textes sous une forme lisible pour l'ordinateur; transcrire les résultats des élaborations avec des machines qui habituellement ne possèdent pas suffisamment de caractères pour couvrir la totalité des divers graphèmes présents dans les textes; mettre au point les programmes et les traduire en langage accepté par l'ordinateur. Le premier problème est peut-être le plus important: les données à perforer et donc à transcrire habituellement sont très nombreuses. Le coût est très élevé, et la série de caractères disponibles sur les perforatrices, et généralement sur les machines pour la préparation de l'*input*, est « absolument insuffisante ». Le chercheur fixe alors des codes plus ou moins complexes pour chaque ponctuation, pour chaque signe diacritique, pour chaque changement de corps, etc. Mais le schéma de codification peut devenir rapidement trop compliqué et souvent il transforme des opérations très simples, comme le *scanning* du texte, comme le classement alphabétique, comme le décompte des fréquences, etc., en opérations beaucoup plus complexes que ce qu'elles pourraient être normalement. Aussi le chercheur évite-t-il le plus grand nombre de distinctions possible. Cela nous est arrivé au début du projet pour le *Dictionnaire historique italien*. D'autres chercheurs ont décidé que les différences entre lettres minuscules et lettres majuscules, entre caractère italique et caractère romain, ou entre divers types de variantes, ou au niveau de la ponctuation même auraient été des « luxes » trop coûteux et n'en ont pas tenu compte par mesure d'économie. Ils se sont aperçus, alors qu'il était déjà trop tard pour revenir en arrière car le travail de perforation était déjà avancé, que de telles informations étaient plus importantes qu'ils ne l'avaient pensé. En outre, la codification doit être établie en parfaite collaboration entre le chercheur linguiste, philologue, humaniste et le programmeur. En réalité les programmeurs et autres spécialistes sont portés à considérer les ordinateurs comme des machines à calculer qui peuvent exécuter d'autres travaux seulement grâce à quelques « trucages » et astuces. Ils ne se rendent pas compte que l'élaboration de données linguistiques est en fait une discipline spécialisée qui a déjà accumulé un nombre considérable d'expériences et de techniques. (D. G. HAYS, 1969, p. 1). Ainsi, soit le chercheur soit le programmeur pensent souvent être en train de faire quelque chose de nouveau, d'original, d'exceptionnel et, en réalité, ils sont en train de perdre du temps et de l'argent pour réinventer la roue. (M. KAY, 1967, p. 171). D'autre part, il existent en beaucoup d'endroits, le fait est plus ou moins connu, des programmes qui font exactement ce que le chercheur demande: concordances, index, recherche de syntagmes, etc. Il semble qu'il suffise de les demander à leur auteur pour les employer. Mais par la suite il y a toujours quelque chose qui ne va pas: souvent on découvre qu'ils exigent une codification un peu diffé-

2. PROJETS DES INSTITUTS USAGERS

Je ne crois pas qu'il soit possible d'énumérer ici un à un tous

---

rente, ou exigent un système électronique un peu différent lui-aussi, et il y a toujours quelques raisons plus ou moins obscures pour lesquelles la personne qui la première a utilisé ces programmes ne veut pas ou ne peut pas les modifier; le programmeur qui les a composés est parti, ou a autre chose à faire; et tout le monde sait par expérience que « toucher » au programme d'un autre n'est pas simple du tout.

A cause du nombre sans cesse croissant des ordinateurs et de la multiplication des projets, d'année en année, il est de plus en plus probable qu'un texte demandé par un chercheur déterminé soit déjà enregistré électroniquement et élaboré.

D'autre part, dans de nombreux pays, y compris l'Italie, certaines maisons d'édition sont en train d'adopter les systèmes de *type-setting* et de *photocomposition* et produisent aussi naturellement, des enregistrements sur fiches ou sur bande de textes publiés. Nous sommes déjà en train de travailler afin que la disponibilité de ce matériel soit assurée aux recherches linguistiques. La perforation des données initiales et la programmation de l'ordinateur pourraient devenir moins coûteuses et moins complexes pour le chercheur individuel. Il s'agit en définitive, étant donnée l'importance de l'effort économique et du temps nécessaire, aussi d'un problème d'ordre moral, disons plutôt, d'éthique professionnelle: dans l'immense majorité des cas, les recherches sont subventionnées entièrement par le Trésor public, et en Italie presque toujours par le C.N.R.

Beaucoup seraient surpris de savoir le nombre de versions mécanographiques différentes qui existent pour un même texte, pour les poèmes d'Homère, ou la Divina Commedia, et le nombre de programmes différents que peuvent donner les mêmes dépouillements et les mêmes statistiques sur le même type de machine.

Il y a aussi une exigence d'ordre scientifique: les résultats du dépouillement des textes particuliers, devraient même pouvoir se prêter à la comparaison avec les autres textes du même auteur, ou avec un nombre considérable d'oeuvres de la même époque, de la même école, ou du même genre littéraire, etc.

Il est nécessaire d'adopter un système *standard* permettant de représenter n'importe quel nouveau texte, et de créer ainsi une grande bibliothèque électronique de textes enregistrés avec les mêmes critères techniques et scientifiques qui puissent être élaborés avec les mêmes programmes fondamentaux.

Il faut une série de programmes généralisés pour la procédure toute entière de dépouillement (lexical, phonétique, syntaxique, etc.) et pour les élaborations statistiques s'y rapportant. Les expériences accumulées de nos jours permettent sans doute de programmer de telles procédures de façon que, au moyen de quelques *control-cards* seulement, elles soient utilisables pour élaborer n'importe quel texte en n'importe quelle langue, ou au moins tous les textes pouvant être ramenés à une écriture de type alphabétique. Il faut que les chercheurs soient rapidement informés sur les textes déjà disponibles ou sur les initiatives et les projets relatifs à des textes nouveaux, de façon à éviter toute répétition inutile.

Il faut surtout des centres spécialisés pour regrouper et mettre à jour les connaissances et les techniques dans le secteur, qui progresse rapidement, de l'élaboration automatique de données linguistiques, et pour éviter aux linguistes et aux humanistes en général de perdre leur temps à étudier des problèmes techniques déjà résolus depuis longtemps et à apprendre des notions et des langages de programmation qui n'ont rien à voir avec leurs intérêts professionnels. Signalons que la section linguistique du CNUCE a effectué ce travail pour l'Italie. Nous avons préparé un système de codification qui est utilisé pour des textes très différents si l'on considère leur genre, leur époque et leur langue si bien qu'aujourd'hui la bibliothèque électronique du CNUCE compte plusieurs

les projets en cours, d'autant plus que la plupart d'entre eux seront

---

dizaines de millions de mots. Le travail du chercheur afin d'établir les normes de perforation ne demande ainsi, au maximum, que deux ou trois heures, au lieu des jours et des semaines nécessaires habituellement. Il ne s'est pas encore trouvé de cas où des textes s'avéraient dépourvus d'informations utiles à quelques nouvelles recherches. Certains ont dit que la standardisation devrait se limiter à l'enregistrement des textes au niveau d'archives électroniques (par exemple à l'enregistrement sur bandes magnétiques) et qu'elle peut être difficilement étendue aux systèmes de perforation. Un système de perforation standard, aussi bon soit-il, pourrait s'avérer trop compliqué pour des problèmes particuliers très simples car étant trop général. Par exemple pourquoi réserver des codes pour les accents graves, aigus et circonflexes si le texte à perforer est en latin? Les codes concernés pourraient être utilisés à d'autres fins. Le système pourrait être amélioré en choisissant les codes les moins coûteux pour les caractères les plus fréquents; mais s'il est vrai que dans n'importe quel texte les lettres minuscules sont plus fréquentes que les lettres majuscules il est vrai aussi que la fréquence de certains graphèmes dépend des langues ou des textes particuliers. Pour cela, certains préfèrent que la standardisation soit effectuée en passant d'un code de perforation, laissé au libre choix du chercheur, à un système de codification standard dans les archives. Notre expérience nous montre qu'un système standard de perforation peut fonctionner dans des conditions presque optimales pour une quantité de textes de loin supérieure à ce que l'on affirme habituellement. Et s'il s'agit de Centres qui travaillent sur différentes langues, il suffit que le système soit optimal pour la langue et le type de textes prédominants: le temps perdu à perforer un texte pour lequel le système n'est pas optimal est le plus souvent inférieur au temps que l'on perdrait à familiariser les opérateurs avec des formes nouvelles et différentes de perforation. Notre programme *cards-to-tape* a une «routine» généralisée de conversion qui répartit les diverses informations contenues dans les fiches en catégories fonctionnelles distinctes. La standardisation devient vraiment possible grâce à une classification des graphèmes et des informations présentes dans les textes, classification effectuée, en prenant comme critère distinctif la fonction, ou plus précisément l'ensemble des fonctions complexes que chaque graphème ou chaque information doit exercer dans les élaborations successives de la procédure toute entière. En ce qui concerne le problème de l'impression, nous avons projeté l'acquisition de quelques chaînes spéciales et nous les avons demandées à l'IBM: en particulier deux *print-chain* de 120 signes destinées respectivement à l'alphabet latin et à l'alphabet grec et un *print-train* de 120 signes pour l'alphabet latin. Nous pouvons ainsi éviter les précautions nécessaires pour résoudre le problème des caractères qui s'avéraient naturellement homographes quand nous disposions seulement de chaînes de 48 caractères. Cet équipement devrait solutionner le problème de l'impression des documents de travail et des résultats à reproduire photostatiquement pour des éditions à tirage réduit (j'ai suggéré à certains instituts qui effectuent leurs dépouillements auprès du CNUCE et qui ne peuvent pas publier tous les index et concordances qu'ils produisent, de les mettre à la disposition des chercheurs en reproduisant photostatiquement sur demande les textes tabulés) ou à reproduire en offset. (C. TAGLIAVINI, 1965; A. ZAMPOLLI, 1967; ACCADEMIA DELLA CRUSCA, 1967; les concordances non lemmatisées de l'*opera omnia* de Sénèque sont en cours d'impression). La tecnique actuelle de la photocomposition est pourtant préférable (cf. D. G. HAYS, 1966, *Introduction*) puisqu'elle joint à la qualité esthétique de l'impression et à la variété presque illimitée des styles, corps et caractères que l'on peut employer dans une même page, l'exactitude absolue de la composition. Tandis que, à cause des erreurs fréquentes dans la chute mécanique des caractères entraînés par la force de gravité, le *type-setting* quant à lui exige la lecture des épreuves. En décembre 1968 nous avons imprimé, en collaboration avec l'Académie de la Crusca, un premier petit volume d'index et de con-

l'objet de communications au cours de ce congrès;[1] il est peut-être plus utile d'indiquer les principaux secteurs d'activité des Instituts intéressés.

2.1. *Documentation lexicographique pour la rédaction des grands dictionnaires historiques d'une langue.*

Exemples: *Dizionario Storico della Lingua Italiana* de l'ACCADEMIA DELLA CRUSCA; *Dizionario di testi storici dell'eteo cuneiforme*, de l'ISTITUTO DI GLOTTOLOGIA DI PAVIA, pour le Centre d'Etudes Egéo-anatoliques du CNR; *Dizionario storico della lingua rumena del XIV secolo*.[2] Toutes les entreprises de cet ordre exigent ordinairement l'élaboration de textes très différents entre eux, du point de vue de la datation, de la langue, du genre littéraire, des critères d'édition. D'une part, les programmes doivent contenir des algorithmes flexibles, adaptables aux structures et aux codifications des textes; d'autre part, puisque les matériaux doivent à la fin se trouver réunis dans les mêmes archives, leur codification doit être uniforme comme dans un corpus unique, pour assurer à celui qui consulte les attestations se rapportant à un même mot, réunies à partir des sources les plus disparates, une compréhension exacte et la plus directe possible des informations délivrées. Pensez à la variété des types de références, à la nécessité d'unifier les marques de l'appareil critique (intégrations, suppressions, lacunes, etc.) et d'unifier aussi généralement tous les différents procédés, adoptés pour représenter au niveau graphémique un même phénomène, entre lesquels hésitent les éditeurs. Pensez surtout aux formula-

---

cordances en photocomposition, qui est la première tentative de ce genre en Italie et nous semble-t-il en Europe.

A la fin de l'année nous publierons les index et les concordances, ainsi que le dictionnaire de rimes, relatifs au *Canzoniere* de Pétrarque).

[1] Les paragraphes suivants (de 2.1 à 2.11) ne doivent absolument pas être considérés comme une classification mais seulement comme une division visant à simplifier notre exposé. On sait que les classifications jusqu'alors proposées de cette ensemble d'activités que les anglo-saxons groupent sous le terme de *Computational Linguistics* n'ont pas donné des résultats pleinement satisfaisants. Je rappelle pour mémoire les classifications de S. LAMB, une dans *Language* en 1961 et une dans le volume *The Use of computer in Anthropology* en 1965, celle de J. GARDIN dans le même volume, celle de P. GARVIN dans *Language* en 1962, celle de C. A. MONTEGOMERY présentée à l'ICCL en 1969 et celle de D. G. HAYS présentée au *Second International Congress of Applied Linguistics* en 1969.

[2] Voir dans ce volume les articles respectifs de A. DURO, P. MERIGGI, et de F. DIMITRESCU-NICOLESCU.

tions du lemme, échelonnées sur plusieurs étapes par différents lemmatisateurs, pour des textes qui appartiennent à des couches de langue différentiées soit synchroniquement (styles, dialectes), soit diachroniquement.

2.2. *Documentation lexicographique destinée à la rédaction de dictionnaires historiques se rapportant à des disciplines particulières.*

A titre d'exemple, je rappelle le projet pour la compilation du *Lessico Intellettuale Europeo*[1] et le projet du *Vocabolario Giuridico della Lingua Italiana*:[2] l'un et l'autre suivent les techniques et les méthodologies mises au point pour l'*Accademia della Crusca*, et se proposent, en particulier, comme résultat final d'obtenir des fiches-contexte qui, en partie du moins, devront se retrouver aussi dans l'*Archivio della Lingua Italiana* d'où sera tiré le *Dizionario della Crusca*. En plus des difficultés qui dérivent de la variété des textes dépouillés, ces projets sont compliqués, du point de vue de l'élaboration, par la nécessité d'effectuer une plus grande sélection dans le dépouillement: en fait le lexique qui devra être mis en évidence dans les oeuvres analysées et transcrit ensuite sur des fiches, est un lexique nettement spécialisé; pour cette raison le rapport en pour-cent entre les fiches-contexte finales et les fiches contenant en perforation la totalité des mots de chaque oeuvre est très bas.

De ce fait, ces projets prennent l'aspect d'expériences économiques intéressantes dont l'importance apparaît clairement si l'on pense à la demande quotidienne de dictionnaires terminologiques, auxquels sont intéressés les chercheurs de nombreuses disciplines scientifiques et de différents secteurs techniques. Cela donne souvent lieu à des projets internationaux en plusieurs langues. Les lexicographes ont souligné combien le lexique se dilate par suite de la pression d'une nouvelle découverte et d'une nouvelle technique. De nombreuses entreprises internationales, conscientes de ce fait, se lancent dans des dépouillements lexicaux automatiques quotidiens de textes techniques pour mettre au point et tenir continuellement à jour inventaires et dictionnaires terminologiques bilingues, imprimés ou bien seulement mis aux

---

[1] Le projet du *Lexique Intellectuel Européen* qu'aurait dû nous présenter ici son directeur malheureusement absent comprend des analyses sur des oeuvres intéressant l'histoire de la pensée, l'histoire des sciences, l'histoire de la culture, dans la mesure où ces analyses ont un rapport avec les buts institutionnels de celles-ci.

[2] Voir l'exposé de C. CIAMPI.

archives et consultables électroniquement, pour être utilisés immédiatement par des traducteurs (humains) de chroniques, de manuels d'enseignement, etc., dans des buts scientifiques et industriels.

2.3. *Documentation lexicographique destinée à la rédaction de lexiques, d'index et de concordances de l'« opera omnia » d'un auteur.*

Toujours à titre d'exemple, je cite les dépouillements de l'« *opera omnia* » de Dante Alighieri, d'Antonio Rosmini, et de L. A. Seneca. *L'Index Thomisticus* a effectué pendant quelques années ses élaborations au CNUCE; il est maintenant transféré aux USA pour quelques années, nous laissant la copie des matériaux (bandes magnétiques). Cela ne nous semble pas être un hasard si dans cette catégorie rentrent surtout des philosophes et des penseurs. L'index du lexique est considéré comme la clé permettant de pénétrer la pensée de l'auteur, comme un moyen de recherche des passages où sont définies ou explicitées les idées. L'étude de l'évolution de l'usage terminologique est révélateur de l'évolution de la pensée; pour tout cela, il convient naturellement de dépouiller l'*opera omnia.*

2.4. *Dépouillements d'oeuvres d'auteurs particuliers, pour différentes études (métriques, grammaticales, stylistiques, thématiques, etc.) effectuées le plus souvent sur une base lexicale et statistique.*

Chaque type d'étude mériterait qu'on lui consacre un discours, que l'on pourrait facilement imaginer: en fait, sont ici présents des experts de chaque section. Pour cela, je me limiterai à citer quelques unes des langues (grec classique, latin-classique, médiéval, de la Renaissance et moderne, espagnol, français, allemand, sanscrit, hébreux, araméen, paléoslave, italien depuis les origines jusqu'à nos jours etc.), et quelques textes (*Tabulae Iguvinae*), la *Gotische Bibel*, le thâatre latin en entier, la législation justinienne, des oeuvres de D. Fabbri, Saba, Machado, Vallejo, Gide, Petronio, Goethe, Pindaro, Bacchilide, etc.).[1]

---

[1] L'expérimentation d'une lexicographie sanscrite mécanisée, en collaboration avec l'*Istituto di Glottologia* de l'Université de Pise, se propose d'obtenir à partir d'un seul input, c'est-à-dire en effectuant seulement une transcription et une perforation, deux formes de transcription: une en *padapatha* et une en *samdhi* sur lesquelles on puisse travailler séparément, soit dans le cadre d'une enquête lexicographique normale soit dans tous les autres domaines de l'enquête statistique et algorithmique. A propos du projet des *Tabulae Iguvinae*, voir A. PROSDOCIMI, 1970; pour le théâtre latin voir la communication de T. GUARDI; pour les textes justiniens, voir la communication de U. ZILLETTI.

Il peut être intéressant de signaler quelques expériences de concordances différentes des formes traditionnelles.

Un premier type est celui que nous appelons *concordances contrastives* qui se sont révélées très utiles pour faciliter la confrontation entre un texte original et les différentes traductions en d'autres langues, ou pour confronter sur un plan critique les rédactions successives d'un texte.

On procède de cette façon. Le texte base (par exemple, l'original ou la rédaction la plus ancienne) est subdivisé en unités contextuelles (péricopes) optimales, dans des limites de longueur convenues, et énumérées progressivement. L'on subdivise les textes « satellites » (par exemple, les traductions ou rédactions postérieures) de la même façon en ayant soin de faire correspondre à chaque péricope du texte original une péricope équivalente dans chaque texte satellite. Les péricopes reçoivent le même nombre progressif, qui permettra à l'ordinateur de jumeler la péricope originale et ses traductions. De cette façon, à chaque forme du texte original on attribue comme contextes non seulement les péricopes dans lesquelles cette forme apparaît, mais aussi les péricopes qui correspondent à ces unités contextuelles dans les différents textes satellites. De la même façon chaque forme des textes satellites reçoit comme contexte, soit les péricopes dans lesquelles elle apparaît, soit les péricopes correspondantes des autres textes satellites et du texte base.[1] Comme on peut facilement l'imaginer, cette méthode qui a permis d'analyser avec soin les diversités lexicales et stylistiques entre les différentes traductions, ouvre des perspectives intéressantes pour les études sur le « mécanisme » de la traduction humaine.

Un deuxième type, qu'on peut appeler *concordances thématiques*, a été requis pour l'analyse de la structure des textes, en particulier des textes poétiques qui se basent sur une tradition orale. Pour simplifier la description, je me réfère à un exemple concret: l'analyse des poésies de Pindare et de Bacchilide. Le chercheur dresse une liste de *thèmes*, à partir desquels la structure des pièces serait composée, et analyse ultérieurement la structure des thèmes en plus petites unités (*motifs*). Par une phase de préédition, le chercheur repère les thèmes qui apparaissent dans la structure d'une pièce et délimite et attribue à chaque thème le morceau de texte (péricope) qui lui correspond. De la

---

[1] Naturellement il peut arriver que dans les textes satellites on ne puisse pas trouver une péricope équivalente à quelques péricopes du texte de base, et vice versa. Le programme de concordances met en relief ces cas particuliers.

*

même façon, il délimite à l'intérieur de chaque péricope les sous-péricopes qui correspondent aux motifs. On obtient ainsi, à côté des concordances lexicales normales, d'autres concordances ou fiches-contexte dans lesquelles l'exposant principal est constitué par les codes qui représentent les thèmes, ou les motifs, ou leurs combinaisons, tandis que les éléments lexicaux y figurent comme des sous-lemmes: les contextes sont constitués par les péricopes relatives. Naturellement, on élabore aussi des différentes listes de fréquence qui mettent en évidence la distribution du lexique en relation avec les thèmes et les motifs, et vice versa. La préédition et la programmation tiennent compte du fait qu'une péricope correspondant à un motif peut-être « encastrée », toute entière ou en partie, à l'intérieur d'une péricope correspondant à un autre motif.

Nous parlerons plus loin au n. 4 d'un autre genre de concordances qu'on appelle *syntaxiques* qui permettent d'une façon encore plus claire ces entrelacements et encastrements de péricopes.

2.5. *Philologie.*

Comme chacun sait, les expériences qui ont été conduites jusqu'à maintenant pour automatiser la *critique des textes*, concernaient surtout la phase de la *collatio codicum*, même si des modèles pour automatiser ou au moins pour orienter la *recensio* ont été proposés. (D. J. FROGER, 1968). Ces expériences ont été l'objet de nombreuses critiques; surtout en ce qui concerne la *collatio*, on a insisté sur le coût élevé de la transcription, de la perforation, des contrôles de tous les codes, qui dans l'ensemble annulerait les avantages de la confrontation automatique.

Il y a cependant d'autres applications dans lesquelles il apparaît évident que l'usage de l'ordinateur apporte des avantages non seulement scientifiques mais aussi d'ordre économique. Un exemple très clair est celui du projet d'élaboration de textes poétiques italiens du treizième siècle, exposé dans la conférence du professeur D'ARCO SILVIO AVALLE. Je me limite ici à quelques considérations d'ordre méthodologique. La procédure projetée prévoit 4 phases: la rédaction du texte critique; la perforation et son contrôle; l'élaboration des concordances, des fréquences lexicales, du dictionnaire des rimes et des incipits, etc.; l'impression automatique (en photo-composition, probablement) du texte, des concordances, des index statistiques et lexicaux.

Avec une seule transcription du texte, et donc par un seul contrôle, on obtient soit le texte critique, soit les index et les concordances.

De plus les corrections exécutées sur une bande magnétique sont plus sûres et plus économiques que les corrections d'épreuves, tant et si bien que de nombreux éditeurs, de périodiques surtout, préfèrent aujourd'hui enregistrer et corriger le texte à publier à l'aide d'un ordinateur, et l'imprimer ensuite directement à l'aide de quelque mécanisme qui ne nécessite pas la correction des épreuves.

Cette méthode permet de publier les textes après avoir élaboré et examiné les index et les concordances, que l'expérience commune fait figurer au rang des instruments les plus efficaces pour la découverte des erreurs casuelles dans le texte; en particulier, pour la mise en évidence des incohérences dans le comportement du philologue (par exemple, en ce qui concerne les variantes de graphie). Les concordances et les index que l'on produit chaque jour au sein du CNUCE ont fait ressortir des erreurs et des coquilles dans des éditions réputées à juste titre comme étant excellentes et dignes de confiance.

Le projet d'analyse critique comparée des 3 rédactions successives de l'*Orlando Furioso* de LUDOVICO ARIOSTO m'amène à parler du problème posé par la façon de traiter les variantes et de traiter, en général, l'appareil critique dans les dépouillements lexicaux. Nous adoptons des techniques de complexité diverse. La solution la plus simple exclut, des concordances, les variantes: elle prévoit uniquement l'apposition d'un signe, en exposant ou même dans le contexte, pour avertir le lecteur que tel ou tel mot a des variantes. Dans les index, au contraire, les listes et les fréquences des variantes constituent des appendices autonomes.

Les solutions les plus complexes exigent, au contraire, l'insertion des matériaux lexicaux contenus dans l'apparat critique, soit dans les index, soit dans les concordances.[1]

---

[1] *Dans les index* apparaîtraient non seulement des fréquences distinctes des formes dans le texte et dans l'appareil mais aussi leurs relations réciproques, c'est-à-dire: pour chaque forme du texte, la liste des formes qui en constituent les variantes et les fréquences relatives; pour chaque variante, la liste de toutes les formes dont elle est la variante et les fréquences relatives. *Dans les concordances* on ferait apparaître en exposant aussi les variantes. Le contexte compterait aussi, grâce à un système opportun de parenthèses, les variantes des mots ou des syntagmes qui le composent. Dans le cas particulier de l'*Orlando Furioso*, on pense à une solution de ce genre: chaque contexte serait composé de trois lignes; dans la première le vers tout entier de la rédaction définitive, dans la deuxième et la troisième ligne seulement les variantes, par rapport à l'édition définitive, de la première et la deuxième rédaction.

2.6. *Dialectologie.*

Quelques recherches récentes, menées indépendamment en Europe et en Amérique, ont démontré que la dialectologie est un domaine de la linguistique particulièrement ouvert à l'utilisation des ordinateurs, pour des buts multiples. Gestion automatique des archives d'éléments dialectaux. Reliefs statistiques: pensez, par exemple, à la liste des oppositions phonématiques. Compilation mécanique de cartes[1] dans la forme traditionnelle, c'est-à-dire relatives à un mot ou à un syntagme du questionnaire des enquêtes; ou bien, relatives à un phénomène linguistique particulier: par exemple, relatives à tous les mots qui commencent par un certain groupe consonantique. Possibilité d'explorer rapidement les archives toutes entières pour extraire des indices régionaux ou locaux. De toute façon, pour mettre en évidence des concomitances géographiques de faits linguistiques définissables dans leur forme.

En collaboration avec le *Laboratoire dialectologique de l'Université de Turin*, nous avons établi, pour des *recherches* en dialectologie diachronique, des concordances constrastives de versions successives de textes dialectaux appartenants à différentes périodes; nous affrontons maintenant le problème de la mécanisation des réponses dialectales au questionnaire de l'*Atlante Linguistico Italiano* (*ALI*) — environ neuf millions de mots en transcription phonétique — pour en tirer ensuite, automatiquement, les cartes dialectales et les index relatifs:[2] le projet est décrit par son directeur C. GRASSI dans ce volume.

---

[1] Je rappelle les travaux de W. N. FRANCIS, J. SVARTIK, G. M. RUBIN de la Brown University sur les matériaux enregistrés dans le quatrième volume du *Survey of English dialects*; les travaux de G. CASSIDY à l'Université de Wisconsin pour le *Dictionary of American Regional English*; les travaux de R. W. SHUY pour l'*automatic retrival* de l'*Atlas Linguistique* des U.S.A. et du Canada, et ceux de G. R. WOOD pour la « découverte » et l'étude des différents types de l'américain régional le long de la côte atlantique.

[2] Les projets pour un atlas dialectologique italien présentés à plusieurs reprises (1909, 1914, 1921) par d'illustres linguistes (D'Ovidio, Goidanich, Parodi, Salvioni, Bartoli, Bertoni, etc.) se concrétisèrent enfin après la première guerre mondiale grâce à la Società Filologica Friulana « G. I. Ascoli » dans le cadre du vaste projet de l'ALI (*Atlante Linguistico Italiano*) dont les enquêtes menées surtout par U. PELLIS en 1.000 points de la péninsule et des îles, et portant sur un questionnaire de plus de 7.000 questions prirent fin en automne 1965 sous la direction du regretté B. TERRACINI. Voir le *Bollettino dell'Atlante Linguistico* publié à Turin. On peut démontrer:

– que notre technique de codification phonétique résoud les problèmes d'input tout comme la chaîne de transcription résoud ceux d'output.

– qu'il est utile de bouleverser l'ordre des phases de la procédure traditionnelle (qui

On pense déjà à un nouveau projet de recueil des matériaux dans toutes les régions d'Italie, et les promoteurs supposent de nouvelles archives dont les articles pourront devenir le *heading* d'une encyclopédie électronique des connaissances dialectales.

A ce propos, on doit tenir compte du renouvellement d'intérêt pour la dialectologie en Italie, celui-ci étant motivé, entre autre, dans le contexte italien actuel de mobilité sociale et géographique de groupes et d'individus, par la position privilégiée des études dialectologiques et régionales dans le cadre des recherches socio-linguistiques.

2.7. *Traditions populaires.*

Ces applications sont effectuées entre autre pour constituer des archives d'informations et des répertoires de musées, ou bien pour faire des recherches de produits d'art populaire: on a inventorié et analysé, en particulier, les structures musicales de chants populaires. Je m'arrête un instant sur ces aspects qui intéressent la *Computational Linguistics* tandis que, en ce qui concerne la description des projets en cours, je vous renvoie à la communication de A. M. CIRESE. Les éléments qui passent d'un texte à l'autre dans les processus de transformation, sont des groupes de mots fixes, des formules cristallisées, qui représentent le véritable « lexique » à l'aide duquel sont composées les poésies populaires. On nous demande si l'élaboration électronique peut contribuer à la détermination de ces formules et, si oui, comment elle peut y contribuer.

---

nécessiterait d'abord la mise sur carte des matériaux, et ensuite leur classement dans un index d'accès aux cartes) dans cet ordre-là:

a) Enregistrer mécanographiquement tous les mots.

b) Dans des buts pratiques d'indexation et partout où cela est possible, « uniformiser », c'est-à-dire, ramener ces mots à une forme unique italienne ou toscane équivalente.

c) Elaborer les archives ainsi obtenues selon les paramètres permis par les diverses catégories d'information enregistrées.

d) Publier, avant même les cartes, les index alphabétiques régionaux, grammaticaux, phonétiques, les tables dialectales comparatives et les autres documentations complètes du matériel rassemblé.

e) Choisir les mots à mettre en carte avec l'avantage méthodologique de dominer, grâce aux index, la masse des données (par exemple repérer des mots dialectaux obtenus pour ainsi dire, par hasard, en réponse à des questions non destinées à les provoquer, etc.).

f) Mettre en input l'enregistrement sur fiche ou sur bande des mots choisis et composer automatiquement en photocomposition la carte (ceci est possible, et j'ai vérifié, en déterminant, sur un système de coordonnées cartésiennes, les points-localités et l'espace disponible pour y imprimer à côté la réponse).

On doit tenir compte qu'il ne s'agit pas toujours de groupes de mots identiques, qui pourraient être facilement repérés en dressant des listes de fréquences et la concordance de mots binaires, ternaires, etc., (voir les articles de P. EVRARD et R. MARTIN). Il s'agit souvent de formules au sens propre du mot: alors que la structure reste invariable, les éléments lexicaux peuvent varier. Même dans ce cas, concordances et dictionnaire d'incipits fournissent une documentation utile pour commencer le répertoire des formules, mais il faudrait automatiser la recherche des formules déjà connues, à l'intérieur des textes. Le cas le plus simple serait celui d'une structure reconnaissable par la succession ordonnée d'éléments lexicaux connus (éléments clé), en établissant éventuellement un nombre limite maximal de mots intercalés entre eux. Le cas le plus complexe se présenterait si l'identification des formules nécessitait la reconnaissance d'une structure syntaxique particulière reliant entre eux les mots-clé. Ou, encore, si les « noeuds » de la structure pouvaient être occupés par des classes particulières de signifiés au lieu de mots-clé bien définis, énumérables de manière explicite. Dans ce cas, se poseraient les problèmes bien connus se rapportant à l'analyse syntaxique automatique et à la structuration de « thesauri » spéciaux, compliqués par la nature poétique, orale et pluri-dialectale des textes. Il va de soi que si jamais ces opérations étaient mécanisées, en plus de l'identification de formules connues, on pourrait se permettre l'utilisation de procédures de découverte fondées, par exemple, sur la fréquence particulière des structures syntaxiques et des « co-occurrences » d'éléments sémantiques particuliers, dans des positions déterminées de la composition.

Les correspondances de rimes sont très souvent irrégulières, en ce sens qu'elles ne respectent pas les règles qui, au contraire, sont normales dans la poésie docte. C'est donc un procédé insuffisant que celui adopté pour les dictionnaires de rimes traditionnels, qui réunissent et organisent les vers en se basant simplement sur l'identité complète des graphèmes compris entre la dernière voyelle tonique du vers et la fin du vers même.

C'est un premier pas que nous pensons avoir fait en compilant des dictionnaires de rimes de type nouveau dans lesquels pour chaque mot ou syntagme final de vers sont énumérés tous les mots, ou les syntagmes qui dans les compositions du corpus sont en fait placés en « position de rime » avec le mot en question.[1]

---

[1] En phase de préédition à chaque composition est attribuée une représentation du schéma métrique, du type *ab ab cc* ... par exemple, ou un classificateur de schéma métrique. (cf. A. ZAMPOLLI, 1967). On constitue ainsi une série de paires de mots dont le

C'est non seulement pour éviter l'alourdissement considérable de la préédition que les chercheurs jugent nécessaire d'automatiser, au moins en partie, le processus de reconnaissance des schémas strophiques, mais encore parce que les répondances phoniques que l'on ressent acceptées par la conscience populaire sont encore, en partie, à rechercher et à inventorier. Pour de telles recherches, il sera probablement utile de faire une transcription phonématique du texte.

Les textes appartiennent à différents dialectes et donc la « lemmatisation » pose de graves problèmes théoriques et pratiques.

Aux homographies que l'on trouve à l'intérieur d'un système, on doit en fait ajouter les homographies entre des formes appartenant à des systèmes différents. On se demande même si l'on doit, ou non, formuler un dictionnaire de lemmes (it. « lemmario ») appartenant à un système linguistique choisi comme référence, dictionnaire auquel on pourrait rattacher, au moyen d'un système de renvois, les dictionnaires des autres systèmes. Les unités lexicales de systèmes linguistiques distincts, même s'ils sont voisins, ne se recouvrent pas toujours entièrement. Mais le fait que l'on doive regrouper sous une même unité signifiant-signifié des unités qui, en plus de signifiants souvent très différents, possèdent des signifiés communs en partie seulement, peut faire surgir de nombreux et délicats problèmes. Ce fait peut compliquer la structure d'un dictionnaire éventuellement destiné aux machines. De semblables difficultés existent du reste dans d'autres projets qui doivent élaborer des corpus pluri-dialectaux comme, par exemple, dans le *Tesoro dell'Accademia della Crusca* ou dans l'édition des compositions poétiques du treizième siècle. Mais dans l'élaboration des textes populaires, le problème présente une composante en plus, puisque, comme on l'a vu, la lemmatisation devrait servir à la découverte de formules, structures dont les éléments sont constitués non tant par des mots que par des classes de signifiés.

2.8. *Psychiatrie, Psychologie, Pédagogie.*

L'étude psychologique du comportement linguistique a reçu une nouvelle impulsion du fait de l'influence « chomskienne ».[1] Les cher-

---

nombre est donné par la formule $\sum_{i=1}^{m} n_i(n_i - 1)$ où m est le nombre des relations de rimes et n le nombre de vers qui entrent dans chaque relation, dans la même composition, selon une rime déterminée.

[1] « Nel campo della psicologia, e in contatto con la linguistica trasformazionale, si è avuto un profondo rinnovamento ». (G. LEPSCHI, 1969, p. 14).

cheurs utilisent en ce domaine le matériel documentaire et les exemples obtenus par suite de nos dépouillements de textes.[1]

Voici la prémisse théorique d'autres recherches que les psychiatres et les psychologues de l'Université de Pise mènent avec notre collaboration sur des enregistrements de conversations (ou, de toute façon, de textes oraux) de malades mentaux ou d'enfants retardés: «même la structure formelle du texte, et non seulement ses aspects thématiques se rapportant au contenu, serait, d'une façon, ou d'une autre, l'expression des caractéristiques psychologiques de celui qui parle».[2]

Les premiers résultats, déjà publiés, ont démontré qu'il existe entre les sujets normaux et les différentes catégories nosologiques (surtout lors de l'utilisation des catégories grammaticales, des adjectifs et des temps futurs des verbes, en particulier) des différences statistiques qui pourraient être utilisées pour le diagnostic et pour une plus profonde compréhension de la maladie mentale. L'analyse morphologico-lexicale du langage d'enfants « hypodoués » intellectuellement, et des enfants de dotation mentale normale, provenant du même milieu socio-économique et culturel a donné des résultats analogues. Evidemment, le *Lessico di frequenza dell'italiano contemporaneo* (LIF; cf. n. 3.3.1) constituera un élément de confrontation et de référence pour les études de ce type. Nous avons déjà pu noter que les catégories grammaticales se regroupent en classes dont les variations de fréquence entre les différentes catégories nosologiques présentent des analogies avec les variations de fréquence dans les sous-ensembles du LIF. Naturellement, les réserves faites plus loin sur les prémisses théoriques de la statistique

---

[1] Il est vrai que dans la conception générativo-transformationnelle de la grammaire le corpus a une valeur différente de celle que lui attribuent les autres écoles linguistiques, par exemple les distributionalistes ou les grammairiens traditionalistes, lorsqu'ils prennent leurs exemples dans la littérature.

Les écoles génératives observent que le corpus, aussi vaste soit-il, est fini par définition, tandis que « les linguistes ont généralement admis qu'une grammaire doit être capable de prévoir, à partir d'un nombre nécessairement limité d'observations, un nombre indéfini de phrases qui ne figurent pas dans ces corpus, et qui toutefois seraient considérées par les locuteurs comme faisant partie de la langue ». (N. RUWET, 1968, p. 36). Donc la présence ou non dans un corpus ne peut être retenue comme critère de « grammaticalité » d'une phrase. D'autre part, un corpus peut comprendre des erreurs d'attention, des lapsus, des mots incomplets, que les locuteurs « refuseraient », corrigeraient. Toutefois « c'est 'l'exécution' qui fournit les données d'observation (corpus de tout genre, écrits ou oraux: conversations enregistrées, interviews, pièces de théâtre, articles de journaux, textes littéraires, etc.), qui permettent d'aborder l'étude des 'compétences' », (N. RUWET, 1968, p. 18).

[2] Voir l'exposé de P. CASTROGIOVANNI.

linguistique sont aussi valables pour ces études, mais sont en partie dépassées par le caractère essentiellement inductif et expérimental de la recherche.

2.9. « *Information retrieval* » *et recherches documentaires.*

Un exemple frappant d'activités qui traditionnellement sont regroupées sous l'étiquette d'*information retrieval* est constitué par les expériences qu'a commencées l'*Istituto per la documentazione giuridica del CNR*, expériences qui, en premier lieu, devront permettre la mise en service au niveau national d'une banque des données juridiques.[1]

Tandis que l'intérêt des linguistes pour le langage naturel est donné par définition, l'intérêt offert par d'autres disciplines dérive d'une attention portée non au langage pour lui-même, mais à la fonction du langage en tant que véhicule primaire pour la transmission des informations dans la société humaine. Que l'on accepte ou non l'idée de la soi-disant « explosion de l'information », c'est un fait établi que de nombreuses organisations occupées à élaborer l'information visent à construire des systèmes automatisés pour analyser le contenu de textes en langage naturel. En théorie, le *content-analysis* consiste à déterminer les « concepts » présents dans les textes et les « relations » entre ces concepts. Les concepts et les relations identifiés sont traduits dans un ensemble de phrases canoniques qui représentent le contenu du document: lorsqu'il s'agit d'un système *fact retrieval or question answering*, ces phrases représentent, pour ainsi dire, la connaissance du système, et servent de base pour produire des réponses pratiques à des demandes spécifiques. Naturellement, les procédures de *content analysis* existant actuellement sont seulement une approximation lointaine de cet idéal. Il est évident que la mécanisation de l'analyse du contenu d'un texte en langage naturel implique un ordre élevé de formalisation des connaissances syntaxiques et sémantiques, qui est un des centres d'intérêt de nombreux courants linguistiques modernes et de grande partie de la *computational linguistics*, que certains de ses théoriciens veulent tout simplement limiter à l'étude de la « computérisation » de telles formalisations.

S'il n'existe pas de modèles d'analyse du contenu fonctionnants réellement et donc s'il n'existe pas de méthodes d'*information retrieval* fondées sur cette analyse, cependant, les techniques et les procédures

---

[1] Voir l'exposé de C. CIAMPI.

qui, déjà, jouent un rôle important pour l'analyse de documents, ne manquent pas. Je me réfère par exemple, à la communication de L. FOSSIER sur l'analyse du contenu de sources diplomatiques du Moyen Age. En ce qui concerne les recherches sur la mécanisation progressive de l'analyse (par exemple pour réduire de plus en plus les interventions manuelles dans l'introduction des descripteurs dans le texte) nous sommes moins pessimistes s'il s'agit de documents du Moyen Age, parce que les structures de tels documents, qu'elles soient linguistiques ou qu'elles se rapportent au contenu, peuvent être ramenées à une typologie connue, en bonne partie, a priori.

Je désire souligner les analogies importantes entre les méthodes illustrées par L. FOSSIER, et les méthodes utilisées par le CNUCE dans des projets analogues en cours d'expérimentation. Je cite à titre d'exemple le projet de l'*Istituto di Storia medioevale e moderna, Paleografia e Diplomatica dell'Università di Pisa*,[1] qui portera sur un corpus médiéval très vaste, qui comprendra aussi des documents inédits, dont le texte perforé permettra l'édition automatique.

Nous sommes en train de mettre au point un système de descripteurs, insérés dans le texte, pour définir les unités référencielles, et un système pour les relier entre elles. La phase de préédition des textes a dans ce projet une grande importance. La division du texte en péricopes selon la tradition diplomatique, contribue à la compilation de concordances particulièrement adaptées aux études diplomatiques, qui du reste recevront des documentations statistiques utiles sur les systèmes de datation, sur la nature juridique des documents, sur les formules, etc. Ces statistiques sont possibles parce que, dans la préédition, on remplit pour chaque document une «matrice» où apparaissent codifiées de nombreuses informations historiques.

2.10. *Histoire de l'Art.*

L'application des ordinateurs électroniques à l'histoire de l'art a pris différentes formes, parmi lesquelles celle qui s'est imposée le plus est, sans doute, l'automation des archives et des répertoires d'objets d'art, particulièrement la gestion des catalogues des musées et sa coordination nationale et internationale.

Les applications des ordinateurs électroniques dans les disciplines archéologiques sont désormais si nombreuses que l'on organise des

---

[1] Voir S. P. P. SCALFATI, 1970.

Congrès Internationaux spécialisés. Dans ces secteurs l'*Istituto di Archeologia dell'Università di Pisa a* projeté une élaboration sur le catalogue de la céramique grecque du D. Beazley, comme premier noyau de fichier-archive de la céramique grecque, pouvant être géré automatiquement et à mettre à jour régulièrement. En plus des gestions normales des archives, cet Institut se propose de relier statistiquement les différentes informations enregistrées pour chaque objet.[1]

2.11. *Musique et Musicologie.*

C'est dans une catégorie tout à fait particulière que nous devons placer les applications de l'ordinateur dans le secteur de la musicologie et de la musique. En ce qui concerne la musicologie, les projets en cours au CNUCE pour l'étude du style des partitions, sont comparables à l'étude du style dans le domaine de la poésie. En ce qui concerne la musique le *Laboratorio Fonologico del Conservatorio di Firenze* est en train d'élaborer des programmes pour l'utilisation de l'ordinateur comme exécuteur de musique, pour enrichir la série déjà importante des instruments électroniques utilisés dans les laboratoires musicaux.[2]

3. PROJETS DE RECHERCHE

La Section Linguistique du CNUCE effectue aussi ses propres recherches, financées en partie par le *Centro Nazionale delle Ricerche*, avec un groupe de chercheurs attitrés et en collaboration avec des groupes d'étude d'autres Universités.

---

[1] Un tableau mis à jour de la situation est fourni par les *Preprints* pour le *Colloque International sur l'emploi des calculateurs en archéologie: problèmes sémiologiques et mathématiques* organisé par le CNRS français à Marseille en avril 1970. Pour le projet de Pise voir l'exposé de P. ARIAS.

[2] Le programme préparé élabore des données relatives à l'émission de 30.000 fréquences audibles dont la production a lieu dans une section de l'unité centrale de l'IBM 7090. Le programme permet non seulement la réalisation d'une quantité très élevée d'intervalles et de rythmes inédits et interdits à n'importe quel autre instrument de musique, mais encore accepte une série d'instructions concernant la modification immédiate de fréquence et de temps de tous les sons, ou d'une partie des sons, d'un texte musical quelconque précédemment exécuté ou même simplement mémorisé dans les mémoires centrales et périphériques.

Enfin, on peut modifier le « tempérament » de chaque texte donné à l'aide d'une seule instruction qui détermine l'entité, presque illimitée, de la variation soit en augmentation soit en diminution, du rapport intervallaire originaire. (P. GROSSI, 1969).

3.1. *Recherches dans le secteur « software ».*

Je me limite à énumérer les principaux arguments pour lesquels on a mis au point, ou on est en train de réaliser, des programmes d'utilité (utility programs):

dépouillements lexicaux: production de listes de lemmes et de formes classées alphabétiquement et à fréquence décroissante, des dictionnaires inverses, des dictionnaires d'incipits et de rimes, des concordances de lemmes et de formes, des fiches-contexte;[1] *consultations de lexiques*

---

[1] Nous entendons par fiche-contexte une fiche mécanographique qui porte imprimées les données présentes sur la fiche lexicographique traditionnelle (lemme ou exposant, auteur, oeuvre, référence, date et un contexte que nous appelons « long » ou « macro-contexte » 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, oeuvre et date). Les critères de délimitation du contexte sont désormais en train de s'uniformiser auprès des centres de lexicographie qui utilisent des équipements mécanographiques. On peut pratiquement les regrouper suivant les types que voici:

a) Dans les systèmes de *Kwic-index*, mis ou point pour les applications documentaires surtout, tous les mots des titres qui composent la liste bibliographique à élaborer (ou, plus souvent, les mots lexicaux seulement), sont classés par ordre alphabétique et reçoivent comme contexte les titres, ou des parties de titre, dans lesquels ils apparaissent. Les programmes de *kwic-index* plus répandus de nos jours ne permettent pas de dresser des concordances utiles à un but lexicographique, pour différentes raisons.

b) Le mot exposant est toujours au centre de son contexte, ou bien il est toujours précédé et suivi d'un nombre égal de frappes ou de mots. Le programme est simple à rédiger et la composition des contextes très rapide; souvent, si l'on choisit cette méthode on décide de ne pas lemmatiser, dans le but évident d'exploiter au maximum la rapidité de la machine et d'éliminer toute intervention humaine (cf. *Concordance to Livy*, 1968).

c) Le contexte est constitué par une unité de référence entière: le vers, le paragraphe, l'alinéa, etc. (cf. C. TAGLIAVINI, 1965).

d) Les limites du contexte sont marquées en phase de « préédition »; le texte est subdivisé en péricopes au moyen de marques qui sont perforées et conservées dans les élaborations successives: chaque péricope fait fonction de contexte pour tous les mots qui la composent.

e) Le contexte est toujours et seulement constitué par 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 pour contexte le syntagme tout entier, c'est-à-dire qu'ils ont le même contexte.

f) Le contexte est choisi en fonction de la nature du mot: pour les mots grammaticaux il s'agit souvent d'un « trinôme » dont le mot grammatical est au centre; pour les prépositions on choisit le plus souvent les deux mots qui suivent, etc. La condition nécessaire doit évidemment être que les mots dont on doit établir le contexte soient déjà classés, d'une façon quelconque; 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, dans la phase de lemmatisation par exemple). A ce propos il est intéressant de considérer les possibilités d'automatiser au moins en partie l'ana-

automatiques: mémorisation et consultation d'un lexique automatique, en portant une attention particulière à l'amélioration des temps de recherche.

---

lyse syntaxique et de choisir ensuite pour chaque mot cette partie finale de la structure dont il fait partie et qu'on juge intéressante comme contexte pour la catégorie grammaticale à laquelle le mot appartient.

g) L'ordinateur fournit, par une méthode quelconque (le plus souvent de type *d*), un premier contexte souvent trop abondant, qui est ensuite réduit aux dimensions exigées au moyen de fiches par lesquelles on communique à l'ordinateur les mots que le chercheur, après un examen approfondi, a décidé de supprimer pour alléger le contexte. (Cf. ACCADEMIA DELLA CRUSCA, 1968).

h) Le contexte est construit en tenant compte de signes déterminés comme la ponctuation, le changement de référence, etc. Comme dans le type *b*, le contexte est établi pour chaque mot, de telle sorte qu'il varie d'un mot à l'autre, mais à la différence du type *b* et à l'image, au contraire, des types *c*, *d*, *e*, il est construit en fonction de la présence d'éléments bien définis, de sorte que le mot peut se trouver placé diversement dans le contexte: au début, au milieu, à la fin, suivant les cas.

L'algorithme de notre programme est virtuellement en mesure 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.

Dans notre programme les éléments du texte sont classés de la manière suivante:

les éléments qui doivent avoir un propre contexte et qui doivent être présents dans les contextes des autres éléments (par exemple les mots « lexicaux »);

les éléments qui doivent faire partie des contextes des autres éléments mais qui ne reçoivent pas de contexte propre (par exemple les signes de ponctuation);

les éléments qui ne doivent ni faire partie d'autres contextes ni recevoir un contexte propre; il s'agit habituellement de codes introduits dans le texte pour réaliser certaines opérations du programme (par exemple les signes de division en pages et en lignes);

les éléments qui doivent recevoir un contexte propre mais qui ne doivent pas entrer dans la composition des autres contextes (par exemple les variantes, les enclitiques, etc.).

En ce qui concerne les éléments du texte on doit faire aussi une seconde distinction selon deux groupes:

les éléments qui ont pour fonction de délimiter le contexte (par exemple un signe de ponctuation ou la fin d'un chapitre);

les 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'au nombre de 99. Les éléments de la classe 99 sont des « limites infranchissables » du contexte; c'est-à-dire que si dans la construction de la partie droite d'un contexte on rencontre une limite de classe 99, le contexte ne pourra en aucune façon s'étendre davantage sur la droite quelque soit l'espace encore disponible. On considère souvent comme limites de classe 99 les marques d'un changement de chapitre, évidemment puisque l'on ne juge pas utiles, pour le contexte du dernier mot d'un chapitre, les premiers mots du chapitre suivant et vice versa.

Les éléments des autres classes, des classes 01 à 88, sont des « limites franchissables » du contexte. Si en construisant la partie droite d'un contexte on rencontre une de ces limites, le contexte pourra la dépasser et continuer son extension vers la droite seulement après avoir trouvé une limite de la même classe ou de classe supérieure dans la progression vers la partie opposée. Les différents signes de ponctuation peuvent être répartis

algorithmes de parsing: pour l'instant nous nous limitons à des algorithmes d'analyse reliés à une grammaire « *context free* » et à l'utilisation sur notre ordinateur, du programme pour vérifier les règles de transformation de *J. Friedman.*

dépouillements phonétiques: algorithmes pour la transcription phonématique semi-automatique de textes, pour la division en syllables, etc.

élaborations statistiques: décompte des différentes unités linguistiques, et calcul de différents indices: dispersion, usage (voir n. 3.3.2.), entropie, etc.

impression avec chaînes spéciales et préparations des données pour la photocomposition.

3.2. *Le Dictionnaire Italien de Machine.*

La Section Linguistique du CNUCE a en projet un *Dizionario Italiano di Macchina* (DMI), dont le noyau comprendra 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, périodiquement avec les résultats des dépouillements effectués auprès du CNUCE. Pour chaque lemme nous enregistrerons, dans une première phase d'élaboration, l'appartenance éventuelle à des secteurs particuliers du lexique,[1] les rapports d'homographie [2] ou de « renvoi »[3]

---

en différentes classes, selon leur valeur: par exemple le point, le point d'interrogation, le point d'exclamation en classe 03; les parenthèses, les guillemets, les deux points, le point virgule en classe 02; le tiret et la virgule en classe 01, etc.

[1] Par exemple les termes techniques, les mots archaïques, les formes dialectales, etc.

[2] 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 différenciés 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*, « franco 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.

[3] Nous désignons par ce nom 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 expo-

à d'autres lemmes, l'étymologie, les affixes et les affixoïdes, la

---

sant *ricuperare*, ou bien avec un exposant *ricuperare, recuperare*. Naturellement ce type de renvoi se différencie en sous-type particuliers selon que les flexions du lemme variant sont distinctes de toutes les flexions du lemme principal (exemple *recuperare, ricuperare*) ou bien coincident en partie (ex. *tonare, tuonare* « tonner ») ou bien que le lemme variant soit défectif (ex. *dicere, dire* « dire »).

Ce schéma devra être complété, si l'on veut organiser dans le DM 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 DM 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, adv. et prép.* 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 plus enclitiques, les prépositions articulées, les mots composés avec trait d'union, les mots avec affissoïdes. 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 « occurences » 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 flexion autonome de l'article; *b*) cumuler les fréquences des différentes prépositions articulées en une forme ' abstraite ' définie comme « forme de l'article lié à des prépositions »; *c*) ajouter ' l'occurence ' 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 DM un quatrième type de « renvois », relatif aux unités de texte composées par plusieurs unités graphiques. Ces unités engendrent des doutes chez le lemmatisateur plus encore pour leur détermination que pour leur traitement; elles exigent en plus que l'algorithme de consultation du DM 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.

transcription phonématique,[1] une première classification grammaticale,[2] et une série de *codes de flexion* qui permettront d'obtenir immé-

---

Le lemmatisateur, pour opérer les choix permis par les différents types de « renvois » devra compiler un petit nombre de *fiches-contrôle* prévues par le programme de consultation du DM, ou bien il pourra attendre que le programme choisisse, comme l'on dit habituellement, par *default*, ces alternatives qui sont adoptées plus fréquemment, selon notre expérience, par les lemmatisateurs.

Pour une définition de préfixoïdes et de suffisoïdes, cf. B. MIGLIORINI, (1941).

[1] L'algorithme qui produit les lemmes et les formes en transcription phonématique exige que dans les *lemmes* soient déjà indiquées la position de l'accent tonique, l'ouverture du *e* et du *o* toniques, la sonorité du *s* et du *z* (quand on ne peut pas les déterminer d'une façon combinatoire), la qualité semi-vocalique du *i* et du *u*. Ces indications sont données suivant le système phonématique florentin savant comme il est témoigné, par P. FIORELLI, dans les mots du *Dizionario Enciclopedico Italiano* et du *Dizionario Zingarelli* (1970) en tenant compte des alternatives proposées de temps en temps (par exemple *microbo : micròbo* « microbe », *altalèna : altaléna* « balançoire », etc.). Les indications susdites ne sont pas ajoutées aux lemmes manuellement mais automatiquement, et puis contrôlées en les confrontant avec les dictionnaires cités, afin de corriger d'éventuelles erreurs faites par l'algorithme qui, comme il n'existe pas de « règles » de correspondance entre les graphèmes et les faits phonétiques susdits, opère sur des bases statistiques avec toutefois de bons résultats: par exemple l'algorithme d'accentuation a donné 5 % d'erreur. Pour les verbes « réguliers » on doit indiquer aussi la voyelle sur laquelle est retiré l'accent dans certaines formes et l'ouverture éventuelle de *e* et *o* qui se trouvent dans cette position (ex. *rotolàre : ròtolo* « rouler, rouleau »; *controllàre : contròllo* « contrôler, contrôle »); pour les « irréguliers » la flexion se produira à partir de plus d'un thème, chacun étant fourni par les indications susdites.

Le système de transcription phonématique laissera aux intéressés le choix entre les différentes alternatives qui font que l'inventaire phonématique est contreversé. (Voir 3.3.2.b); celui-ci fournira aussi les éléments nécessaire pour tenir compte au niveau syntagmatique des phénomènes de phonosyntaxe: par exemple on marquera toutes les formes qui provoquent le renforcement de la consonne initiale suivante.

[2] Voici une première liste de classificateurs grammaticaux:

I interjections.
CC conjonctions de coordination.
C autres conjonctions.
LP locution prépositive.
LA locution adverbiale.
L autres locutions.
BS adverbes non qualificatifs.
B autres adverbes.
N numéraux (cardinaux).
SM substantif masculin.
SF substantif féminin.
SN substantif mobile (avec subdivision ultérieure relative au traitement différent dans les dictionnaires: par exemple masculin et féminin constitués en lemmes autonomes, ou réunis sous le lemme du masculin, etc.).
A adjectifs (ni pronominaux ni numéraux).
AN adjectifs numéraux.
PP pronom possessif.
PQ » personnel.

diatement la flexion automatique.[1] L'algorithme de flexion produira les formes, soit en graphie normale, soit en transcription phonématique et nous aurons ainsi, à côté d'une liste de lemmes, aussi un dictionnaire de formes; chaque forme pourra être ramenée automatiquement au lemme dont elle provient, et sera accompagnée aussi par des classificateurs morphologiques traditionnels.[2]

Dans une seconde phase, nous enrichirons la série des informations en ajoutant à chaque lemme des classificateurs syntaxiques,[3] sémanti-

---

PD	pronom		démonstratif.
PI	»		indéfini.
PT	»		interrogatif.
PR	»		relatif.
PE	»		exclamatif.
PF	»		réfléchi.
PC	»		réciproque.
PN	»		numéral.
DP	adjectif	pronominal	possessif.
DQ	»	»	personal.
DD	»	»	démonstratif.
DI	»	»	indéfini.
DT	»	»	interrogatif.
DR	»	»	relatif.
DE	»	»	exclamatif.
DF	»	»	réfléchi.
DC	»	»	réciproque.
DN	»	»	numéral.
B	non codifiés pour incertitudes.		
RI	article indéterminé.		
RD	article déterminé.		
	préposition propre.		
	préposition impropre.		
VT	verbe transitif.		
VI	verbe intransitif.		
VP	verbe pronominal.		
VR	verbe réfléchi.		

[1] Voir pour cet algorithme le rapport de G. FERRARI.

[2] Par exemple le genre, le nombre, le degré pour les noms et adjectifs; la forme, le mode, le temps, la personne pour les verbes.

[3] En effet, nous nous proposons ensuite de commencer les recherches pour un système d'analyse syntaxique automatique de l'italien, et cela pour deux raisons.

Premièrement, à un niveau pratique, nous voulons automatiser autant que possible la séparation des homographes. En italien les homographies entre les lemmes appartenant à des parties différentes du discours sont très fréquentes, plus de 90 % de l'homographie dans son ensemble (par ex. le type verbe-substantif: *faccia* « la face » ou « que je fasse »; *porta* « la porte » ou « il porte »). Un *parser* syntaxique automatique pourrait les résoudre, en distinguant aussi certaines homographies à l'intérieur d'une même partie du discours: par exemple la distinction serait faite non seulement entre *danno* substantif et verbe – mais encore, grâce à l'« accord » avec le sujet, entre *danno* 1ére

*

ques,[1] et statistiques.[2]

Les objectifs du projet DMI peuvent être résumés en trois groupes.

---

pers. du sing. du présent de *dannare* – « je damne » – et *danno* 3ème pers. du pl. de *dare* – « ils donnent » –.

Deuxièmement, nous voulons un instrument qui nous aide dans ce travail complexe que représente la description d'un langage, ou d'une partie de langage, au moyen d'un système formel. On peut prendre pour exemple dans ce domaine le système mis au point par J. FRIEDMAN et ses collaborateurs à l'Université de Michigan et de Standford:

« Writing a transformational grammar for even a fragment of a natural language is a task of a high order of complexity. Not only must the individual rules of the grammar perform as intended in isolation, but the rules must work correctly together in order to produce the desired results. The details of grammar-writing are likely to be regarded as secondary by the linguist, who is most concerned with what is in the language and how it is generated, and would generally prefer to pay less attention to formal and notational detail. It is thus natural to ask if a computer can be used to assist the linguist in developing a grammar. The model is formal: there are a large number of details to be worked out; the rules interact with one another in ways which may not be foreseen. Most of the errors which occur in writing grammars can be corrected if only they are brought to the attention of the linguist. Those which cannot be so corrected should be of even greater interest to the writer of the grammar ». (J. FRIEDMAN, 1969).

D'autre part les études les plus récentes ont toujours accordé la plus grande importance à la fonction du lexique dans les grammaires générativo-transformationnelles. La classification toujours plus détaillée du lexique est aujourd'hui une des voies principales pour le progrès de ces études, et nous pensons que le DM est un instrument indispensable.

Voir les rapports de M. GROSS, M. SALKOFF, P. VEILLON, et la bibliographie citée plus loin à la note 42. Pour un *parser* avec grammaire *contex-free* de l'italien voir le rapport de A. VON STECHOW.

[1] Même si la sémantique, après avoir été pendant longtemps le parent pauvre des sciences linguistiques (T. TODOROV, 1966, p. 5), est aujourd'hui à l'ordre du jour (N. RUWET, 1964), les diverses méthodes (H. RUBESTEIN, 1966) d'analyse jusqu'à aujourd'hui proposées sont bien loin de pouvoir être appliquées à un lexique de dimensions même très inférieures au nôtre.

Les analyses qui nous sont connues se limitent à quelques dizaines de lemmes en général appartenant à des groupes fortement structurés. On ne sait pas encore clairement si cela dépend d'une impossibilité *a priori* (Cf. T. DE MAURO, 1970), ou d'un manque de recherches applicatives (T. TODOROV, 1966, pp. 17-18), de sorte que, comme le prévoit D. G. HAYS (1969), les recherches sémantiques intégrées dans l'étude d'un modèle global de la compétence du locuteur-auditeur constitueront l'objet principal de la linguistique dans les prochaines vingt années. Je crois que la route à suivre est celle indiquée par de nombreux linguistes, par M. ALINEI (1970), par exemple, pour l'italien, c'est-à-dire une recherche progressive visant à extraire des traits distinctifs ultérieurs par des « ricerche strutturali e tassonomiche sui diversi campi semantici che risultano dall'incrociarsi dei tratti combinatori ».

De toute façon, un DM est utile pour enregistrer les résultats de cette recherche, qui demande du temps et la collaboration indispensable de nombreux groupes, et pour pouvoir vérifier à travers la production et le parsing automatique de phrases le caractère adéquat de l'analyse et de ses niveaux (G. J. HUTCHINS, 1971).

[2] Consulter plus loin le § 3.2.3.

3.2.1. *Lemmatisation semi-automatique de textes italiens.*

L'ensemble des opérations regroupées ordinairement sous le nom de *lemmatisation*, nécessite une série d'interventions humaines qui, en brisant le rythme des élaborations entièrement automatiques des dépouillements, en augmentent considérablement le temps et le coût.

La procédure normale de lemmatisation à l'aide d'un DM tend à réduire durée et dépenses; l'ordinateur attribue automatiquement le lemme aux mots qui sont enregistrés comme étant univoques dans le DM, tandis qu'il soumet à l'examen du lemmatisateur les contextes des mots qui ne figurent pas dans le DM, ou bien y sont rapportés comme homographes.[1] Le risque de cette méthode est qu'il est possible d'insérer dans le DM comme univoque une forme qui au contraire est en fait homographe. Dans ce cas l'ordinateur lemmatiserait naturellement les mots correspondants comme univoques sans les soumettre à l'examen du linguiste. Plus la couche de langue examinée s'étend diachroniquemente, plus l'on retourne en arrière dans le temps vers des systèmes qui ne sont plus tout à fait présents à la conscience linguistique de celui qui compose le DM, plus le risque, dont nous avons parlé, est grand. Aussi ce risque a-t-il souvent dissuadé les lexicographes de préparer un DM pour leurs dépouillements.

Cependant, dans ces cas-là, l'obstacle est surmontable si au lieu de considérer comme étant définitive la lemmatisation des mots univoques

---

[1] Nous prévoyons deux méthodes de consultation du DM.

Dans la première méthode, l'élément de recherche sera constitué par les formes, et, par conséquent, il faudra que le texte à lemmatiser soit classé alphabétiquement. Comme l'observe D. G. HAYS (1967, p. 87), il est vrai que l'opération de *sort* est généralement longue et coûteuse mais on doit considérer que la distinction manuelle des homographes et l'éventuel contrôle des lemmes attribués aux formes univoques nécessitent de par eux-mêmes l'examen des contextes classés alphabétiquement.

Dans la seconde méthode, la recherche se déroulera selon le processus bien connu qui prévoit la segmentation du mot en thèmes (pris dans le sens de séquence de graphèmes qui demeure invariable dans la flexion) affixes, et désinences. Donc le DM sera organisé sous forme de liste de thèmes enregistrés sur disque (avec le système des *letter tables* – cf. D. G. HAYS, 1967, p. 92 – étant le nombre réduit – environ 200.000 – par rapport au DM de formes – environ 2.000.000 – qu'il faut inévitablement enregistrer sur bande). Cette liste sera complétée par des tableaux d'affixes et de désinences enregistrés dans la mémoire centrale, et enfin par des codes de « compatibilité » entre les thèmes et les éléments des tableaux.

Dans ce cas, une variante dans la procédure de lemmatisation sera aussi possible, puisque les mots du texte pourront être présents en *input* sans contexte et dans l'ordre du texte. Chaque fois qu'un mot se révèle *homographe*, il peut apparaître sur un écran vidéo avec un vaste contexte et avec l'indication des lemmes possibles, parmi lesquels le lemmatisateur opérerait immédiatement son choix au moyen du *light-pen*.

selon le DM, on imprime leurs concordances lemmatisées afin que le lexicographe contrôle les lemmes attribués par le DM. Les expériences faites à dessein ont démontré que, même avec cette phase de contrôle, le temps épargné serait très important (75 % environ) par rapport à une procédure entièrement artisanale considérée dans ses trois phases: écrire les lemmes à la main, les perforer et les reporter sur bande magnétique, contrôler les deux opérations précédentes.

En procédant de la sorte, je pense que le DM pourra être utilisé aussi dans les dépouillements pour le *Tesoro* et pour le *Vocabolario de l'Accademia della Crusca*, surtout lorsque nous aurons intégré soit le dictionnaire des lemmes, soit l'algorithme de flexion à des éléments et des règles déduits des textes des origines.[1] Autre fait favorable à l'utilisation du DM: celui-ci, en évitant un grand nombre d'inscriptions de lemmes nécessaires à la procédure artisanale, diminue de manière considérable le risque des erreurs casuelles de transcription, et, en même temps, assure une plus grande systématicité 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 DM 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 choisi pour les traiter.[2] En outre, le DM permet la rétrodatation automatique des formes dans le corpus choisi.

Le DM favorise généralement la normalisation dans la lemmatisation des textes, et plus le nombre des textes soumis à dépouillement augmente, plus cette normalisation apparaît actuellement indispensa-

---

[1] Pour des raisons techniques, la consultation du DMI, dans la procédure des dépouillements pour le *Trésor des origines* doit être effectuée par ordre alphabétique. Donc, le DMI pourra être intégré en incorporant les formes repérées dans les oeuvres élaborées et lemmatisées successivement par l'*Accademia della Crusca*, accompagnées naturellement de la date convenue ou connue de l'oeuvre. Cela nous éviterait de formuler l'ensemble des règles permettant de produire automatiquement toutes les flexions et les nombreuses variantes, souvent imprévisibles, rencontrées dans les textes des origines; l'on doit considérer que ceux-ci représentent en réalité des systèmes linguistiques différents, ou des systèmes orthographiques divers, ou non encore établis. En plus de l'économie de temps au niveau de la lemmatisation le DMI permettrait d'autres résultats utiles, parmi lesquels je signale: un répertoire des formes et un répertoire des lemmes continuellement tenus à jour; l'antidatation automatique des mots attestés; le dictionnaire des fréquences; la liste complète des « normalisations » décidée pour les variantes.

J'espère vivement que l'Académie acceptera cette suggestion.

[2] L'importance de cette cohérence est proportionnelle au temps de dépouillement et au nombre de lemmatisateurs qui y travaillent.

ble. Par exemple, les fréquences lexicales de textes divers peuvent être comparées utilement entre eux seulement si ces textes ont été lemmatisés avec des critères semblables ou au moins des critères tels que les résultats du dépouillement puissent être rendus homogènes facilement. 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.[1] L'utilisation du DM garantit que chacun de ces cas reçoit le même traitement de lemmatisateurs différents. Par contre, les formes pouvant être mises sous 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 respectivemment, 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 DMI pourra être adapté au moyen de fiches contrôle de façon à choisir l'alternative préférée pour chaque classe. L'important est que, grâce au DMI, 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.[2]

---

[1] 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 dont j'ai connaissance 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.

[2] En ce qui concerne les effets des critères de lemmatisation sur la validité des dépouillements pour les statistiques lexicales, cf. CH. MULLER (1963) et (1968), A. ZAMPOLLI (1969), et P. GUIRAUD (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 comptés séparément des adjectifs correspondants dans le dépouillement du texte *A*, mais non dans celui du texte *B*, on exclue 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 DMI.

3.2.2. *L'automatisation de la transcription phonématique des textes italiens.*

La transcription phonématique de chaque forme étant donnée dans le DMI, il est possible, par un algorithme de consultation tout à fait semblable à celui de lemmatisation, d'attribuer à chaque mot d'un texte sa transcription phonématique. Naturellement, le chercheur devra examiner seulement les contextes des formes homographes non homophones (par exemple, *áncora – ancóra*; *pésca – pèsca*) pour choisir parmi les différentes transcriptions proposées par le DM. On doit remarquer que pour chaque forme sont aussi reportés des classificateurs qui permettront de tenir compte dans le texte de quelques phénomènes de phonosyntaxe: par exemple, de tenir compte des enclitiques et des proclitiques, du renforcement phonosyntaxique, des nasales finales homorganiques, etc.

3.2.3. *Recherches linguistiques qui présupposent un inventaire des unités du système linguistique italien.*

La linguistique contemporaine est caractérisée par une tendance à formuler des règles exactes et explicites et à énumérer des ensembles complets d'unité linguistique 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 wen 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. FILLMORE, 1968).

On compte déjà quelques exemples de lexique automatique, même établis dans cette perspective.[1] 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 the Theory of Syntax*,

---

[1] Par exemple pour le hongrois le DM de L. PAPP (196), pour le tchèque le DM de J. STÍNDLOVA (196 et 197); 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 DM 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.

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, qui du reste est traité dans les exposés de A. PUGLIELLI, D. PARISI, N. SALKOFF, M. GROSS. 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 DM.[1]

Le DM permet en outre de mener des recherches statistiques sur le lexique considéré comme faisant partie du système linguistique.[2] Ces

---

[1] 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 correspondent chacun à une classe distributionnelle des structures. Une étude générale de toutes les formes verbales est en cours au *Laboratoire d'Automatique Documentaire et Linguistique de CNRS*: 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 DM. Bien qu'ils soient encadrés dans une théorie différente, les articles du DM 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 DM 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 DM. 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 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).

[2] Entre les textes enregistrés sur bande et le DM le rapport est dans une certaine mesure analogue au rapport *langue – parole*. Pour les problèmes relatifs au statut du lexique

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 textes, mais à cause de très grandes difficultés que l'on peut surmonter seulement à l'aide d'un DM.

Les problèmes sont nombreux, qu'ils soient théoriques ou pratiques. Le DM énumère soit les morphèmes,[1] 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 définitions de telles unités, qui soient acceptées universellement et qui soient entièrement satisfaisantes sur le plan opérationnel.[2]

Mais une autre série de problèmes apparaît.

Alors que le vocabulaire, c'est à dire l'ensemble des mots en fait observés dans un texte ou dans 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.

On se pose aussi le problème suivant, souvent débattu mais pas

---

dans la *langue* voir J. RAY-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.

[1] J'emploie ici *morphème* au sens qu'ont donné à ce terme les linguistes américains. (Cf. F. P. HAMP, 1966, pp. 40-41).

[2] Un compte rendu critique de la notion de morphème dans la linguistique moderne a été fait par M. 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. (Par exemple J. H. GREENBERG, 1952, pp. 27-31). Pour les problèmes théoriques de la définition voir par exemple K. TOGEBY, (1949), A. ROSETTI, (1965). Pour les problèmes de la définition opérationnelle voir l'article de CH. MULLER, (1963). Pour la relation entre *morphème* et *mot*, voir, par exemple, J. LYONS, (1968, § 5.3.1-5.4.12), et en particulier, pour l'école transformationaliste, R. P. BOTHA, (1968, pp. 80 et suivantes).

toujours très clairement: le lexique d'une langue doit-il être considéré comme fini ou comme infini?[1].

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[2] communicatives de la communauté parlante.[3]

Un dictionnaire de lemmes, ou de toute façon une nomenclature, ne peut pas donc être considéré de manière absolue comme étant exactement l'équivalent du lexique d'une langue, même si on le juge fini.

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 coincide 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 le lexicographe qui le considère comme un modèle de comportement,[4] et les linguistes, spécialement les générativistes, qui le considèrent comme un modèle théorique.[5] 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 selon des critères

---

[1] 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 d'avantage 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 strictement 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, pp. 116 et suivantes) et par J. RAY-DEBOVE (1970, pp. 3 et suivantes) qui différencie nettement la compétence syntaxique de la compétence lexicale, parce que « si, pour la grammaire, la compétence abstraite et la compétence de quelqu'un coincident grosso modo, pour le lexique il y a un abîme entre la première et la seconde ».

[2] Cf. CH. MULLER, 1968, p. 138.

[3] Cf. L. HIELMSLEV, 1967, p. 97.

[4] Cf. J. RAY-DEBOVE, 1970, p. 7.

[5] Cf. N. CHOMSKY, 1965, p. 3.

les plus explicites possible. Par exemple, sur l'axe diachronique seront marqués dans le DMI les *archaïsmes*, dont la définition est « per l'italiano più difficile che per altre lingue » (B. MIGLIORINI, 1951, p. 18), et dont il existe de différentes sous-espèces et non pas un seul type. (A. J. GREIMAS, 1960). Sur l'axe synchronique on marquera avant tout les vocables dialectaux et l'on marquera à part même «le parole e le locuzioni toscane, e particolarmente fiorentine, che siano ignote all'uso delle altre parlate italiane ». (B. MIGLIORINI, 1951, p. 13). L'on revient ici au problème de l'existence et de la consistence d'un italien *standard*, qui est déjà apparu plusieurs fois dans cette conversation. Sans aucun doute l'état actuel des études de géographie linguistique et de socio-linguistique en Italie ne fournit pas une documentation suffisante pour permettre de tracer avec certitude les limites entre les mots uniquement dialectaux et le lexique commun. Ici aussi les difficultés sont plus grandes pour l'italien que pour les autres langues: « Per molteplici cause (. . .) — l'origine e il carattere spiccatamente letterario della lingua normale, la scarsa coesione sociale e politica degli italiani durante i secoli — l'uso linguistico nazionale è scarsamente compatto ». (B. MIGLIORINI, 1951, p. 6). Il faudra aussi une marque pour les vocables scientifiques et techniques. Ici le problème principal sera la distinction entre les termes utilisés seulement par le spécialiste, etles termes techniques et scientifiques propres aussi au « locuteur idéal », cultivé, mais pas 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 niveasu tylistique 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 la partie centrale du lexique? En tous cas, le décompte statistique des différentes unités sera exécuté séparément pour le lexique « neutre » et pour chacune des catégories marquées. Après avoir examiné les listes des vocables attribuées à chaque catégorie pour vérifier la cohérence de cette attribution, et après avoir comparé les résultats des décomptes des différentes catégories, nous évaluerons diversement les données globales et les données obtenues suivant les différentes combinaisons des sous-ensembles. 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.[1] Comme le

---

[1] 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 LIF nous avons retenu 5.365 lemmes (du reste c'est l'ordre de grandeur pour

reconnaissent les lexicographes, le seul moyen pour accroître le degré d'objectivité ce serait d'avoir recours à la fréquence relative des vocables dans les sous-ensembles d'un corpus représentatif. Ces données étant souvent insuffisantes, il faudra utiliser le DM comme un 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 socio-linguistique, 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 DMI.

Trubetzkoy lui-même donne des exemples sur la façon de procéder et énumérer quelques difficultés de méthodologie.[1] Mais ces difficultés ne suffisent pas à expliquer à elles seules pourquoi le programme des recherches de statistique phonologique tracé par Trubetzkoy et par les premiers pragois a été remarquablement réalisé dans de nombreuses langues[2] 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.[3]

---

de nombreux dictionnaires de fréquences) sur les 15.000 environ donnés par le dépouillement, et ce dernier chiffre aussi est très inférieur au total des lemmes qui constituent le premier noyau de notre DM, environ 120.000.

Les lexicographes se plaignent aujourd'hui de cet état des choses. Par exemple, parlant des limites des applications de la linguistique à la lexicographie, J. RAY-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 ». (J. RAY-DEBOVE, 1969, p. 372).

[1] Cf. N. TRUBETZKOY, 1939, chap. 7.

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

[3] N. SARAMANDU fait allusion de manière explicite à ce chapitre de N. TRUBETZKOY dans son travail sur le rendement fonctionnel des types de structures phonématiques.

Considérons à titre d'exemple en quelles opérations consiste 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 différentiation des signifiants de mots dans une langue donnée ».[1] L'étude de cette portée fonctionnelle, qui intéressa jusqu'à nos jours de nombreux linguistes,[2] nécessiterait en fait:

le choix d'un dictionnaire de lemmes qui représenterait le lexique de la langue en question;

le dégagement de toutes les formes et de toutes les flexions des lemmes variables;

la transcription phonématique de ces vocables;

la subdivision éventuelle des vocables en catégories (par exemple, vocables communs, archaïsmes, dialectismes, etc.);

la découverte de toutes les *paires minimales*, et leur assemblage en autant de groupes qu'il y a de paires de phonèmes distinctifs; [3]

Mais sur les 50.000 lemmes environ du *Dicționarul limbii române moderne* (Bucarest, 1958) l'on à examiné seulement les 2903 lemmes formés de 1 a 4 phonèmes et leurs formes flexionelles ne comportant pas plus de 4 phonèmes (1814).

Le travail de S. GOLOPENȚIA-ERETESCU (1966) se rapporte aussi à des matériaux tramités par le dictionnaire roumain déjà cité et étudie la structure phonologique des monosyllabes qui y sont rapportés.

[1] Cf. J. VACHEK-J. DUBSKÝ, 1960, p. 66.

[2] TRUBETZKOY, 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 of Phonology*, (1955) 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 ».

[3] Un compte rendu des différentes propositions concernantes la définition de *paire minimale* et la bibliographie relative se trouvent dans Ž. MULJAČIĆ (1969, pp. 200 et suivantes, et pour l'italien pp. 406 et suivantes).

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 ». (Ž. MULJAČIĆ, 1969, p. 211). Dans notre DM il serait possible d'effectuer la recherche des paires minimales des morphèmes. En fait, la plus grande partie des structuralistes travaillent sur les

la subdivision des paires minimales de chaque groupe suivant les catégories auxquelles appartiennent les 2 vocables de la paire (par exemple, sous-groupe des paires minimales dont les 2 membres sont des vocables communs, ou bien où l'un des vocables est un vocable commun et l'autre un dialectisme, etc.);

le décompte des paires minimales découvertes pour chaque groupe

---

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, recourent à des paires « dans lesquelles l'un des membres, ou les deux, sont des éléments plus ou moins étendus qu'un simple mot » (G. LEPSCHY, 1964, p. 54). Ce type de paire pourra être étudié lorsque nous pourrons appliquer aux textes le DMI 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-1* 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-1*) 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 ennième 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, § 3.7): en italien par exemple l'accent sur le même emplacement. 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 esemple 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 » et « effrénée » – parce que leurs contextes seraient respectivement différents: *sx-re-n'a-ta* et *sxre-n'a-ta* (x représente la place du phonéme retranché). La situation à ce propos n'est pas très claire. Par exemple 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* a *sfrenata* per fondare una opposizione /e/:/f/ sarebbe molto artificioso, dato che la struttura sillabica è nei due cosí 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 Ž. MULJAČIĆ, (1969, p. 226).

En travaillant sur le DM, il sera facile d'effectuer les décomptes suivant les différentes alternatives. En général, l'emploi du DMI permet une grande flexibilité: par exemple compter ou non les occurrences d'oppositions du type *phonème : zéro* (ex. *dritto-ritto*).

(fréquence absolue de l'opposition distinctive) et pour chaque sous-groupe;

le calcul de quelques fréquences relatives des oppositions: par exemple, leurs différentes positions dans la syllabe ou dans le mot, la fréquence des phonèmes ou des traits pertinents, etc.

Ž. MULJAČIĆ (1969, p. 307): « Per ora è praticamente impossibile trovare tutte le coppie minime in cui possano ricorrere i fonemi di una lingua, soprattutto i fonemi frequenti. Ciò sarà 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,[1] la recherche des paires n'a jamais été faite, même pas pour d'autres langues:[1] nous croyons qu'il est possibile d'essayer seulement avec l'ordinateur, en utilisant un DM 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 leurs combinaisons, et à des données sur les fréquences des mots.[2]

---

[1] « L'attenzione dei primi linguisti statistici che volevano calcolare il numero assoluto di opposizioni basate su un fonema veniva 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 ». (Ž. MULJAČIĆ, 1969, p. 306).

[2] 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, § 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. N. Trubetzkoy avait déjà mis ce point en relief dans le chapitre 7 des *Grundzüge*: « Nicht alle Teile des Wortschatzes sind gleiche und vergleichbar. Es gibt technische Ausdrücke, die nur einem engen Kreise von Fachleuten geläufig sind, obgleich sie keine Fremdwörter im gewöhnlichen Sinne sind. Sollen solche Ausdrücke in die Statistik miteinbezogen werden? Es gibt Wörter, die in ihrer schriftsprachlichen Gestalt wohl nur in Wörterbüchern stehen, faktisch aber nur in dialektischer Lautgestalt leben. weil sie eben ihrem Sinne nach zum Bereiche der Dialektsprachen gehören (vers-

Ce type d'intégration est aussi nécessaire pour les recherches statistiques dans d'autres secteurs. Aussi convient-il d'effectuer une lemmatisation plus « complète » et attribuer à la forme, en plus du

---

chiedene technische Ausdrücke des bäuerlichen Lebens u. dgl.). In welcher Lautgestalt sollen sie statistisch verwertet werden? Solche Probleme tauchen bei der lexikalischen Statistik fast jeder Sprache auf. « Ž. MULYAČIĆ (1969, p. 309) écrit lui aussi: « tutte le coppie minime (...) non hanno la stessa probabilità di comparire nel messaggio. Alcuni loro membri sono arcaismi o tecnicismi e di quelli che non sono rari soltanto una parte può trovarsi nello stesso contesto. Rarissime sono le coppie omofunzionali del tipo *blond-blanc* »

Une technique pour combiner les deux données a été proposée. Si les deux membres de la paire minimale sont de grande fréquence, le rendement de l'opposition dans les *textes* est élevée. Si, au contraire, un des membres de l'opposition est de grande fréquence et l'autre de basse fréquence, alors le rendement de l'opposition dans les textes est faible, même si le nombre de paires minimales dans le lexique est le même, même si les fréquences d'utilisation des deux phonèmes en opposition sont semblables.

L'on propose donc de multiplier, pour chaque paire minimale, la fréquence d'un membre par la fréquence de l'autre, et d'additionner ensemble les résultats ainsi obtenus pour toutes les paires minimales relatives à la même opposition distinctive. Mais il faut ajouter une autre considération: même si les deux membres d'une paire minimale sont fréquents cela ne veut pas dire qu'ils soient en fait opposés fréquemment dans l'usage. Les deux mots peuvent appartenir à des contextes complètement différents de telle sorte qu'ils ne peuvent commuter qu'en des cas extrêmement rares, ou même se trouver en distribution complémentaire.

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 la technique que nous avons exposée, 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 *l'Introduction to Theoretical 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: « To take the limiting case: if A and B are two classes of words in complementary distribution, and if each member of A differs in its substantial realization from a member of B solely in that it has the element /a/ where the corresponding word of B has /b/, it is clear that the functional load of the contrast between /a/ and /b/ is nil. So the functional load of a particular contrast must be calculated for words which have the same or an overlapping distribution. It is also clear that any ' realistic ' measure of the importance of a particular contrast should take account, not simply of the distribution of the words as specified by the rules of the grammar, but of the actual utterances in which the words might be confused if the contrast were not maintained. How often, for instance, or in what circumstances, would an utterance suche as *You'd better get a cab* be confused with *You'd better get a cap*, if the speaker failed to make the distinction in the final consonants of *cab* and *cap*? The answer to this question is evidently relevant to any precise quantification of the functional load of the contrast in question ». Les difficultés viennent du fait que « the linguistics theory, at the present time at least, is not, and cannot, be concerned with the production and understanding of utterances in their actual situation of use ». Il faut donc tenir compte que la théorie concerne uniquement « the structure of sentences considered in abstraction from the situation in whic actual utterances occur ».

En ce sens il est possible d'adopter un procédé de type hiérarchique dans lequel sont importantes pour le calcul non pas toutes les oppositions, sans distinction de position dans

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.

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,[1] et celles de J. DUBOIS sur la relation entre fréquence

---

le mot, mais seulement celles entre phonèmes qui apparaîssent dans la même position syllabique. Le calcul des fréquences des phonèmes dans les différentes positions syllabiques (prénucléaire, nucléaire, postnucléaire) peut être fait selon la technique de KUČERA et MONROE (1968).

La possibilité qu'offre le DM de compléter sur la totalité du lexique la recherche statistique commencée par J. KRÁMSKÝ et portant sur 8.804 mots italiens de1 ,2 et 3 syllabes, pris de dictionnaires bilingues, mérite qu'on y fasse allusion: dans son ensemble le travail parait réalisable seulement au moyen de l'élaboration électronique. J. KRÁMSKÝ s'est proposé, par une analyse quantitative, de « penetrate into the structure of a language. It is the ascertainment of the functional load of a particular liguistic feature that provides the basis for a hierarchic classification of these phonemes. However we cannot be satisfied with a mere statement of the functional load of phonemes or of phonological features. We must try to find out the means which the language makes use of in the most fundamental process of forming word units, that is combining phonemes in units possessing a certain meaning ». (J. KRÁMSKÝ, 1964, p. 129). Pour faire l'analyse « of mutual relations between the phonemes of word units », après avoir réparti les 8.804 mots en 46 types différents (per example *ba*, *ába*, *ábba*, *abába*, *ábaba*, etc., où *a* représente une voyelle et *b* une consonne) il a établi pour chaque type de mot: « 1. The frequency of individual vocalic and consonantal phonemes. – 2. The frequency of phonemes in the interrelation of all positions. Thus, for example, in the word type b1a2a2b3a3x it was ascertained how many, let us say p's ther are in position b2 (or b3) in relation to perticular phonemes of position b1, that is the relations b1-b2, b1-3, and b2-b3 had to be ascertained. In a similar way the relations between phonemes of positions b1-a1, a1-b2, a1-a2, a1-b3, a1-a3, a2-b3, and a2-a3 were examined- 3. The quantitative analysis of consonantal phonemes is doneon the basis of grouping them (a) according to the manner of articulation (plosives and affricates: p, b, t, d, k, g, ts, ds, t, dz; fricatives: f, v, s.; nasals and liquids: m, n. n, 1, 1, r), (b) according to the place of articulation (labials and labio-dentals: p, b, f, v. m; alveolars and apicals: t, d, ts, ds, s, n, 1, r; palatals: t, dez, n, 1; velars: k, g.) ». (p. 130).

Il est évident que le nombre des corrélations augmente rapidement, si l'on étend l'examen à des mots ayant un plus grand nombre de syllabes, et à un lexique plus représentatif que celui choisi par l'auteur, de telle sorte qu'il est indispensable de travailler sur un DM.

1 Les premières études de P. GUIRAUD sur le dictionnaire, par exemple sur la relation entre le nombre de phonèmes d'un mot et ses signifiés et 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, la distribution des dérivés sémantiques et morphologiques en les calculant dans les dictionnaires anglais, français et allemand.

et étymologie, et sur la productivité des systèmes de dérivation du français.[1]

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 constituantes 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 DM à deux niveaux: morphèmes et mots.

3.3. *La Statistique Linguistique sur des Textes.*

La bibliothèque électronique établie auprès du CNUCE, se prête naturellement, comme nous l'avons dit, à des recherches de statistique linguistique et, de façon particulière, à la vérification, sur une base très vaste de données, des hypothèses formulées.

Nous avons engagé les études statistiques à trois niveaux: lexical, phonématique et syntaxique.

3.3.1. *Statistiques Lexicales*

En 1971 sera publié par l'IBM ITALIA et par les soins de U. BORTOLINI, C. TAGLIAVINI et A. ZAMPOLLI un *Lessico di Frequenza della Lingua Italiana Contemporanea* (nous écrirons ici LIF).

---

[1] 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).

*

L'italien a été un peu la langue oubliée en ce qui concerne les dictionnaires des fréquences, qui sont déjà disponibles pour beaucoup de langages.[1] Comme on le sait, les recherches tournées vers la compilation de ces dictionnaires ont commencé déjà à la fin du XIX siècle, et se sont en suite multipliées par un mouvement qui atteint son apogée dans les années qui précèdent immédiatement la deuxième guerre mondiale et qui s'harmonisa avec le développement général de la *linguistique appliquée*, expression qui, surtout aux Etats-Unis, fut pendant longtemps synonyme de *glottodidactie*.

La diffusion, parmi les humanistes, des techniques d'élaboration des données, donna une impulsion nouvelle, dans les années '50, aux projets de dictionnaires fondamentaux, qui purent bénéficier de nouvelles possibilités de dépouillement et d'élaboration statistique des données. Au cours du *Premier Congrès International de Linguistique Appliquée* (Nancy, 1964) on réserva une place de grande importance à la section des dictionnaires fondamentaux. Cette deuxième phase de l'histoire des dictionnaires fondamentaux est caractérisée par une plus grande attention aux problèmes méthodologiques et théoriques. Certes on ne doute pas de l'utilité didactique de listes de mots, particulièrement fréquents dans la

[1] Voir pour la bibliographie P. GUIRAUD (1954) et la série des dictionnaires de fréquence des langues romanes de A. JUILLAND.

Pour l'italien de brefs essais seulement ont été faits jusqu'à présent, celui de T. M. KNEASE, *An Italian word List from Literary Sources*, Toronto, 1933 et celui de B. MIGLIORINI, *Der Grundlegende Worschatz des Italienischen*, Marburg, 1943. Le premier essai représente un dépouillement manuel, sur un total de 400.000 mots environ, de sources exclusivement littéraires comprenant des oeuvres en prose et en vers qui vont de la deuxième moitié du XIXème siècle aux premières décennies du XXème siècle. Presque tous les mots grammaticaux et aussi les autres mots considérés comme étant « particulièrement fréquents » ont été omis, et les homographes n'ont pas été différenciés. Le travail de B. MIGLIORINI a été réalisé uniquement sur la base du sentiment linguistique de l'auteur, illustre historien de la langue italienne et lexicographe expérimenté, qui se proposait de donner à des fins exclusivement didactiques une liste de 1.500 mots qu'il considérait fondamentaux dans la langue italienne contemporaine et, pour cette raison, plus utiles pour les étrangers qui étudient notre langue. Il distingue les 500 mots les plus importants en les imprimant en caractère gras tandis qu'il laisse en caractère normal les autres 1.000 mots. Une confrontation rapide avec notre liste nous a montré que, en plus de certains mots évidemment liés à la situation politico-sociale du moment: *balilla*, *corporativo*, *corporazione*, *aviere*, *duce* (terminologie fasciste) qui figuraient dans le MIGLIORINI parmi les 500 premiers mots et qui ne trouvent pas place dans notre liste, il y en a d'autres dans MIGLIORINI, (par exemple *eroine* féminin d'*eroe* « héros », *voluttà* « volupté », *solere* « avoir coutume », etc.) qui reflètent des aspects lexicaux et stylistiques désormais en décadence. D'autres mots au contraire placés par Migliorini parmi les 1.500, apparaissent dans notre recueil au-dessous du seuil d'usage que nous avons choisi comme limite de notre liste, par exemple: *ardore* « *ardeur* », *artigliere* « artilleur », *attirare* « attirer », *balzare* « sauter », *berretto* « casquette », *fisionomia* « physionomie », etc., jusqu'à un total de 77 lemmes.

langue à enseigner. Mais on veut rendre explicites les prémisses statistiques et linguistiques de cette première intuition, de façon à pouvoir utiliser les résultats et les techniques récemment développés par la statistique linguistique, pour affiner et pour augmenter la rentabilité des procédures de compilation des Dictionnaires de Fréquence.

Les problèmes de l'échantillonnage.

Dans les intentions des compilateurs des dictionnaires de fréquence les textes soumis au dépouillement font fonction d'échantillons de la langue conçue comme un univers statistique. Ainsi se posent les problèmes de l'échantillonnage, pour lesquels la statistique a élaboré toute une série de méthodologies et de techniques (M. BOLDRINI, 1968, pp. 512 et suiv.).

Il faut, d'une part, délimiter l'univers, de l'autre, s'assurer du caractère représentatif de l'échantillon. Si l'on connaît bien les caractéristiques de « l'univers », on peut créer un échantillon dans lequel chacun de ces caractères est distribué selon les mêmes proportions avec une stratification soignée[1] (CH. MULLER, 1968, p. 15). On comprend facilement que ceci n'est pas le cas de la langue.

Les fréquences sont observées dans un texte, c'est-à-dire dans une réalisation du système linguistique de la part d'un individu, réalisation déterminée par ses caractéristiques personnelles d'une part, et, de l'autre, par les caractéristiques de la communauté à laquelle il appartient, caractérisée sur le plan social, géographique, historique, etc. Si le texte est un texte littéraire, même les règles et les particularités du genre littéraire, et dans chaque cas la situation dans laquelle il est créé, ont une influence sur la production de ce texte. Les lexicographes ont toujours parlé d'un langage technique, d'un langage poétique, d'un langage populaire, d'un langage savant,[2] etc. Il convient de se demander si actuellement nous possédons sur la structure de l'univers linguistique une connaissance suffisante pour décider les dimensions et la composition d'un échantillon représentatif, et même pour établir si et dans quelles limites le rapport entre *textes* et *langue* est légitimement assimilable, avec rigueur méthodologique, au rapport échantillon-univers.

---

[1] C'est le cas, par exemple, des sondages d'opinion avant les élections. On connaît avec précision la répartition géographique, la condition socio-professionnelle, l'âge, le sexe, etc., des inscrits sur les listes électorales.

[2] Voir par exemple B. MIGLIORINI (1951, pp. 10 et suivantes et pp. 42 et suivantes).

Il y a au moins deux manières différentes de considérer la langue comme univers statistique.

Selon une première conception, les unités du système linguistique seraient caractérisées non seulement par des traits qualitatifs émergeant des oppositions et des relations qui forment la structure du système lui-même, mais aussi par leurs probabilités d'usage respectives. Ces probabilités ne sont pas directement observables comme ne l'est pas, du reste, le système. Cependant, elles seraient traduites par le fait (tenu pour certain) que les unités linguistiques se répéteraient dans les textes parlés et écrits à des fréquences relativement stables. Dans cette perspective les fréquences observables dans les textes sont considérées comme des « approximations » des probabilités non observables du système.

Cette conception a reçu une formulation théorique explicitée par P. GUIRAUD,[1] même si celle-ci transparaissait déjà d'études précédentes,[2] et fut acceptée et développée en premier lieu par G. HERDAN [3] et ensuite par d'autres chercheurs qui reconnaissent en elle le fondement théorique sur lequel la statistique linguistique pose sa propre autonomie en tant que science.[4]

---

[1] P. Guiraud dit expressément que la langue pourrait être considérée comme un ordre systématique d'« engrammes »; chaque engramme présent avec sa probabilité qui constituerait un attribut objectif comme la forme ou le signifié (P. GUIRAUD, 1960, p. 16 et p. 25). Il a aussi formulé les bases théoriques de la possibilité d'analyser la « langue » et la « parole », assimilées à un « code « et à un « message », avec les concepts et les techniques de la théorie de l'information. Plusieurs aspects du fonctionnement de la « langue » sur lesquels Saussure fonde sa doctrine, comme « la différentialité, la segmentation, la linéarité, l'arbitraire du signe linguistique ... sont ... des caractères généraux et pré-linguistiques communs à toute catégorie de signes informationnels » (P. GUIRAUD, 1954, p. 119).

L. Heilmann reprend le problème en ces termes: « Se, dunque, oggi vogliamo tener presenti tutti i fattori che si riconoscono nel funzionamento della lingua, noi dobbiamo considerare quest'ultima come un sistema qualitativo e quantitativo di segni arbitrari doppiamente articolati (sul piano del significante e su quello del significato), caratterizzati da forma, contenuto e frequenza, attuantisi in un processo individuale fisiopsichico sottoposto al determinismo statistico » (L. HEILMANN, 1963, p. 37).

[2] D. W. Reed, parmi les premiers, dans un article paru dans *Word* en 1949, énonce le concept de fréquence comme caractère constitutif d'une forme linguistique et insiste sur le rapport entre la *fréquence* dans des textes particuliers et la *probabilité* dans la langue (R. W. REED, 1949, p. 235-236).

[3] Herdan propose l'assimilation entre l'antinomie de Saussure « langue-parole » d'une part et le rapport « population statistique-échantillon » de l'autre. La « langue » équivaut à la « totality of engrams in the brains of members of the speech community together with their probabilities of occurence » et la « parole » « for random samples from it » (G. HERDAN, 1960, p. 34).

[4] « Ca și alte ramuri ale lingvisticii moderne, lingvistica statistică – în calitate de disciplină constituită – se reclamă ca avînd bazele teoretice în conceptia despre limbă a lui F. de Saussure ». (A. ROCERIC, 1968, p. 18).

Ces prémisses n'ont cependant pas trouvé, jusqu'à nos jours au moins, un consentement général.[1]

Déjà les auteurs du *Français fondamental*, en introduisant la notion de disponibilité, furent les premiers à ébranler cette conception. R. MOREAU, au congrès de Strasbourg de 1964 sur le thème *Statistique et analyse linguistique*, traçait de manière explicite cette situation en affirmant: « les premiers pas de la statistique appliquée à la linguistique ont précisément consisté à admettre des règles du jeu qui soient simples. C'est ainsi qu'on a énoncé (voir par exemple Guiraud) que la fréquence des mots était constante dans la langue, ce qui supposait donc que l'on pouvait assimiler le choix d'un mot au tirage d'une boule dans une urne dont la composition reste inchangée au cours du temps ». (R. MOREAU, 1966, p. 130). M^me^ L. HIRSCHBERG et CH. MULLER ont démontré aussi que cette régularité n'est pas vérifiée: les fréquences des mots, sauf dans des cas exceptionels, ne sont pas stables mais varient d'un texte à l'autre, sous la dépendance du style et du thème,[2] et cela pourrait être vérifié même « pour *les mots de relation*, qui sont les éléments les moins thématiques du lexique, car beaucoup d'entre eux, la plupart même, subissent visiblement l'effet des situations stylistiques ». (CH. MULLER, 1968, p. 141).

Cela dit en passant, les données que nous avons obtenues dans les dépouillements pour le lexique de fréquence de l'italien, confirment l'existence des variations dans la fréquence des *mots de relation* entre les cinq sous-ensembles examinés, variations qui apparaissent en rapport étroit avec la nature du sous-ensemble, du moment qu'elles ont la même direction pour la plus grande partie des unités appartenant à la même catégorie grammaticale.

Dans la deuxième conception, le concept de probabilité, comme caractéristique intrinsèque de l'unité du système linguistique, n'est pas exprimé; l'univers statistique est défini plutôt comme l'ensemble de tous les textes parlés ou écrits produits pendant une certaine période, Même en admettant que l'ensemble soit exactement définissable, une

---

[1] Quelques auteurs, non seulement des linguistes mais aussi des spécialistes de statistique, ont relevé une excessive simplification dans les affirmations de HERDAN et du « premier » GUIRAUD, et ont proposé une plus grande prudence, en conseillant d'utiliser et, s'il le faut d'attendre, les résultats des dépouillements des textes que le développement de la « linguistique computationnelle » rend toujours plus nombreux et dignes de confiance.

[2] A ce propos, sont intéressantes les considérations de CH. MULLER (1968, pp. 136-141) sur les rapports entre lexique et vocabulaire.

grande partie de cet ensemble échappe à toute mesure: par exemple, les lettres, les conversations privées, etc.

MOREAU distingue deux catégories de mots: les mots de *classe ouverte* et les mots de *classe fermée*.[1]

Les mots de *classe fermée* seraient tous les mots « athématiques », c'est-à-dire, ceux qui nous servent à nous « exprimer à propos des choses plutôt qu'exprimer les choses elles-même ». On pourrait placer dans cette catégorie un certain nombre d'adjectifs et de vocables courants, de mots grammaticaux, et quelques noms très généraux: « des termes plus ou moins communs à tous les sujets et à toutes les situations ». Leur emploi ne semblant pas varier sensiblement dans les différents centres d'intérêt, l'estimation de leur fréquence, à partir de textes échantillon, ne poserait pas de problèmes.

Le cas des mots de *classe ouverte* ou « thématiques » qui, au contraire, présentent oscillations de fréquence d'un texte à l'autre et, souvent, entre les parties d'un texte, est différent. Pour évaluer la fréquence de ces mots, le seul moyen qu'il conviendrait d'adopter consisterait à stratifier a priori la langue, à délimiter des centres d'intérêt, à l'intérieur desquels les mots seraient « thématiques ».

L'idée de MOREAU est précisément que les mots d'usage non stable dans la langue peuvent, au contraire, l'être dans des textes ayant un thème, un centre d'intérêt commun. Il propose donc de stratifier une fois pour toute la langue « à une époque donnée » en un certain nombre de centres d'intérêt, qui pourront être même très nombreux; par exemple: mathématique, physique, chimie, etc. A l'intérieur de chaque couche on déterminera la fréquence des mots thématiques.[2]

---

[1] R. MOREAU, 1962.

[2] Par exemple, dans le centre d'intérêt « algèbre élémentaire », des mots thématiques comme *égalité*, *inégalité*, *équation*, *inconnue*, *variable*, *fonction*, seront « fréquents » et « stables ». Les résultats obtenus sur les textes accessibles à l'échantillonage pourront être étendus à d'autres textes, à la partie inexplorée du centre d'intérêt. Comme il est normal de classer suivant la fréquence les mots généralement stables dans la langue, ainsi pourra-t-on mettre par ordre de fréquence, à l'intérieur de leur centre d'intérêt, les mots thématiques. Mais il y a d'importantes difficultés dues à la détermination des centres d'intérêt. En premier lieu, on se demande si on ne doit pas chercher pour chaque mot, même de fréquence faible, les centres d'intérêt auquel il se rapporte, créant ainsi à la fin un nombre extraordinaire de catégories. Ou s'il ne convient pas, au contraire, de définir a priori, dans la langue, un nombre raisonnable de centres d'intérêt, et de chercher ensuite, à l'intérieur de chacun d'eux seulement la fréquence de quelques mots. D'après Moreau, la réponse se trouve dans la constatation selon laquelle, pour certains mots, il est impossible de définir un centre d'intérêt dans lequel ils reviennent à une fréquence stable. Ce seraient approximativement les mots que le CREDIF définit par l'expression « mots disponibles ». Définir une fréquence pour les mots disponibles n'a aucun sens, écrit Moreau: en effet, le CREDIF a préféré ne pas en indiquer la fréquence dans le *Français fondamental*.

La détermination des centres d'intérêt aura évidemment quelque chose d'arbitraire et la variété de structures et de partitions que nous avons rencontré dans l'échantillonnage des Dictionnaires de fréquence, confirment ce fait.[1] Ceux qui étudient le problème conseillent d'augmenter le plus possible le nombre des couches: « stratifier à outrance », recommande MOREAU (1962, p. 157).

D'autre part, 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 lexicale.

Nous considérons que les études traditionelles de stylistique et de lexicographie peuvent fournir les données de départ pour une toute première stratification a priori de la langue, ou, encore mieux, les données permettant d'identifier, tout au moins comme hypothèse de travail, quelques sous-ensembles suffisemment définis et caractérisés de la même façon par rapport à un paramètre commun. On appliquera ensuite, aux résultats des dépouillements, les techniques statistiques pour vérifier, d'une part, l'homogénéité des structures quantitatives des sous-groupes, et, d'autre part, pour identifier les éventuels mots thématiques.

La reconnaissance effective des couches de langue devra dériver d'un travail d'induction soigné et documenté à partir de dépouillements de plus en plus nombreux, à condition naturellement que ce même travail prouve l'existence, sur le plan des structures quantitatives, de sous-ensembles linguistiques définissables avec certitude. Cette attitude qui est de plus en plus répandue, est du reste favorisée par le développement de la lexicographie automatique dans de nombreux pays, où le nombre de dépouillements disponibles s'accroît chaque 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

---

[1] Les auteurs des dictionnaires de fréquence ont presque toujours composé le corpus à dépouiller avec des morceaux choisis dans des textes divers, définis par une différence d'origine (auteur ou sujet parlant) ou autre (genre littéraire, milieu social, situation concrète, etc.). Non seulement les dimensions globales de l'échantillon choisi pour représenter une langue déterminée sont très diverses, mais encore le choix même des textes répond à des critères différents. De nombreux efforts sont faits pour reproduire dans la composition du corpus échantillon la complexité des catégories et des styles de la langue dans son ensemble. On a des classifications et des subdivisions particulièrement complexes.

italienne comme univers lexical, autour de laquelle on pourra organiser, dans le même but, l'élaboration des données provenant des dépouillements désormais très nombreux qui enrichissent progressivement la bibliothèque électronique du CNUCE.

Au cours de ces dernières années, deux importants dictionnaires de fréquence ont vu le jour (A. JUILLAND, E. CHANG RODRIGUEZ, *Frequency Dictionary of Spanish Words*, The Hague, 1964 et A. JUILLAND, P. M. H. EDWARDS, I. JUILLAND, *Frequency Dictionary of Rumanian Words*, The Hague, 1965), les premiers d'une collection, *The Romance Languages and their Structures* dirigée par A. JUILLAND, qui se propose l'étude quantitative du lexique de la grammaire et de la phonématique des principales langues romanes.

Bien qu'ils aient été sujet à de nombreuses critiques, ce qui d'autre part est inévitable, car ce sont deux travaux nouveaux d'un point de vue méthodologique, ces deux volumes représentent ce qu'on a fait de mieux jusqu'à maintenant dans ce domaine.

Fondés pour chaque langue sur des vastes dépouillements pour un complexe d'environ 500.000 mots subdivisés en cinq groupes différents, ces volumes sont dédiés spécialement à la langue moderne et contemporaine. Nous avons choisi, nous aussi, ce nombre pour le LIF, dans le but surtout d'obtenir des comparaisons plus faciles non seulement avec les volumes rédigés pour deux langues apparentées génétiquement à l'italien comme l'espagnol et le roumain, mais surtout avec le volume qui paraîtra bientôt consacré à l'italien, qui, vue la diversité du corpus échantillon, ne sera certainement pas une répétition de notre lexique.

La principale différence entre notre corpus et les dictionnaires déjà publiés par JUILLAND consiste probablement dans le fait que les textes que nous avons dépouillés (1947-1968) sont plus récents et syncroniques. Aussi la subdivision de l'échantillon est diverse: nous avons remplacé la catégorie *essais* de JUILLAND, qui comprend des essais d'histoire et de critique littéraire, par la catégorie *cinéma*, c'est-à-dire des textes dialogués de film. En outre, on doit remarquer que nous avons pris, au lieu des *textes techniques*, les manuels des écoles élémentaires.

Comme « terminus a quo » nous avons choisi 1945, c'est-à-dire la fin du deuxième conflit mondial (en réalité le moins récent des textes dépouillés a été édité en 1947, le plus récent en 1968). Cette date, bien que ne représentant pas une limite nette, comme du reste les autres fixées pour les périodes historiques ou littéraires, est significative. Elle représente la fin d'une période historique, qui a eu des répercus-

sions importantes sur la langue et marque le début d'une orientation différente dans les sources d'information. Avant tout, un important renouvellement du lexique se produit lorsque deviennent hors d'usage beaucoup de vocables liés à des institutions historico-politiques du régime passé. Deuxièmement, l'influence des modèles français vient à disparaître ou à diminuer considérablement, laissant la place à des modèles le plus souvent anglo-saxons (anglais et surtout américains).

Les 500.000 occurrences de notre échantillon proviennent en parties égales (100.000) des cinq groupes de textes: *Théâtre*, *Romans*, *Cinéma*, *Périodiques*, *Manuels*.[1]

---

[1] Les morceaux ont été pris dans les textes suivants:

Théâtre.

V. BOMPIANI, *Teresa Angelica*, in « Il teatro italiano del dopoguerra », a cura di V. Pandolfi, Bologna 1956.
L. SQUARZINA, *L'esposizione universale*, in « Il teatro italiano del dopoguerra », a cura di V. Pandolfi, Bologna 1956.
C. TERRON, *Avevo più stima dell'idrogeno*, in « Il teatro italiano del dopoguerra », a cura di V. Pandolfi, Bologna 1956.
G. CASSIERI, *Il salto mortale*, in « Sipario », Marzo 1962.
C. FRUTTERO, *Una donna uccise per dedizione*, in « Sipario », Agosto-Settembre 1963.
G. PATRONI GRIFFI, *D'amore si muore*, in « Teatro », Milano 1965.
F. BRUSATI, *La pietà di Novembre*, in « Sipario », Maggio 1966.
D. FO, *Gli arcangeli non giocano a flipper*, Torino 1966.
M. S. CODECASA, *La gara*, Rai-Tv.
M. FIOCCO e M. VERGOZ, *Una mattina d'estate*, Rai-Tv.

Romans.

V. PRATOLINI, *Cronache di poveri amanti*, Firenze 1947.
C. CASSOLA, *La ragazza di Bube*, Torino 1960.
A. MORAVIA, *La noia*, Milano 1960.
F. VITTORINI, *Il Sempione strizza l'occhio al Fréjus*, Milano 1962.
D. BUZZATI, *Un amore*, Milano 1963.
E. ROVERSI, *Registrazione di eventi*, Milano 1964.
G. BASSANI, *Il giardino dei Finzi-Contini*, Torino 1966.
I. CALVINO, *Le cosmicomiche*, Torino 1966.
C. PAVESE, *La luna e i falò*, Torino 1967.
A. BEVILACQUA, *L'occhio del gatto*, Milano 1968.

Cinéma.

P. GERMI, *L'uomo di paglia*, Rocca San Casciano 1958.
R. ROSSELLINI, *Era notte a Roma*, Rocca San Casciano 1960.
L. VISCONTI, *Rocco e i suoi fratelli*, Rocca San Casciano 1960.
DE SICA-FELLINI-MONICELLI-VISCONTI, *Boccaccio 70*, Rocca San Casciano 1960.
M. ANTONIONI, *Deserto Rosso*, Rocca San Casciano 1964.
F. FELLINI, 8 ½, Rocca San Casciano 1965.
V. DE SETA, *Un uomo a metà*, Rocca San Casciano 1966.
M. BELLOCCHIO, *La Cina è vicina*, Rocca San Casciano 1967.

Périodiques.

Quattrosoldi, Agosto 1967.
Corriere della Sera, 15 Maggio 1968.

Comme nous l'avons dit, sauf dans des cas exceptionnels, les fréquences des mots ne sont pas stables, mais varient, d'un texte à l'autre, sous la dépendance du style et du thème. Cette constatation fait naître « l'idée de compléter la fréquence, ou, si l'on préfère, de corriger la fréquence par la façon dont celle-ci se distribue dans le corpus » (CH. MULLER, 1965, p. 34).

Après que le dépouillement lexical ait fourni la liste des lemmes présents dans les textes du corpus choisi comme échantillon, les auteurs procèdent normalement à une sélection et incluent dans le lexique de fréquence seulement une partie des lemmes trouvés. En exécutant un tel choix et en ordonnant les lemmes sélectionnés, les auteurs prennent en considération non seulement la fréquence de chaque lemme, mais aussi la manière plus ou moins uniforme selon laquelle cette fréquence est répartie entre les différents sous-ensembles du corpus.

Parmi les différents indices statistiques suggérés nous avons choisi et adopté dans le LIF la formule de A. JUILLAND et E. CHANG RODRIGUEZ, pour garantir la possibilité de confrontation entre nos données et les leurs.

Comme on le sait, ces auteurs attribuent à chaque lemme un coefficient d'*usage* (U) qui est le produit de la fréquence complexive (F) du mot pour sa dispersion (D): $U = FD$

D est un coefficient qui qualifie l'uniformité de la répartition de la fréquence d'un lemme dans le corpus, s'accroissant linéairement au degré d'uniformité:[1] D varie entre *zéro* (fréquence accumulée en un seul

---

Corriere della Sera, 26 Maggio 1968.
Oggi, 6 Giugno 1968.
Corriere della Sera, 28 Giugno 1968.
Corriere della Sera, 9 Luglio 1968.

Manuels d'école.

M. G. GABRIELI, *Vele*, sussidario per la 3ª classe elementare, Bergamo 1967.
M. G. GABRIELI, *Vele*, sussidario per la 4ª classe elementare, Bergamo 1967.
M. MALFATTI, C. NOTA, U. PETRINI, *Primato*, sussidario per la 5ª classe elementare, Bergamo 1967.

[1] Le procédé pour le calcul de la dispersion, que nous exposons ici avec le système de symboles proposé par CH. MULLER, (1965) dans le compte rendu sur le *Frequency Dictionary of Spanish Words*, exige de calculer avant tout:

$S^2 = \frac{1}{n} \sum_{i=1}^{n} (f_i - f)^2$ où n est le nombre des sous-ensembles, f la fréquence moyenne du mots dans le corpus, et f_i sa fréquence dans le enième sous-ensemble.

On calcule ensuite:

$V = \frac{S}{f}$, où $S = \sqrt{S^2}$

sous-ensemble) et *1* (répartition parfaitement uniforme de la fréquence dans les n sous-ensembles: $f^1 = f^2 = \ldots = f^n = F/n$).

Le LIF contiendra les lemmes (5.356) qui ont $U \geqslant 1{,}78$ dans le corpus dépouillé.[1] Nous les énumérerons par ordre alphabétique, chacun étant suivi par toutes ses formes apparaissant dans le corpus. Lemmes et formes seront accompagnés par la fréquence complexive respective (F), par les fréquences dans chacun des 5 sous-ensembles (fi), par l'usage (U) et la dispersion (D).

Tous les 5.356 lemmes seront aussi énumérés suivant l'ordre décroissant de F, de U, et de D. Pour des motifs d'espace évidents nous n'imprimerons pas dans ces 3 ordres les 30.000 formes, environ, du corpus, mais seulement les 9.000 premières formes, environ, de chaque ordre.

La dispersion est:

$$D = 1 - \frac{V}{\sqrt{n-1}}$$

De très nombreuses discussions ont mis en évidence les qualités et les défauts de cet indice (cf. le compte rendu de MULLER) et certaines modifications ont été proposées: voir les articles de S. ALLEN, J. B. CARROLL (1970), et de J. MISTRIK (1969).

Nous avons préféré la formule de JUILLAND et de CHANG RODRIGUEZ sans modifications, parce que, comme nous l'avons déjà dit plusieurs fois, il nous semble important, dans l'état actuel des recherches de statistique linguistique, de permettre la comparaison entre les résultats des différentes recherches.

[1] Nous avons abandonné la méthode de JUILLAND et de CHANG RODRIGUEZ en ce qui concerne le choix de la limite inférieure d'acceptation des lemmes de notre lexique de fréquence. En effet ils ont choisi les matériaux à publier dans une liste de 9.000 lemmes environ, restés après l'élimination, en un premier temps, de tous les lemmes ayant $F < 4$ (6.000 environ) et puis de ceux ayant $R < 3$ (c'est-à-dire présents seulement dans un ou deux sous-ensenbles: 5.000 environ). Ils expliquent ces décisions par des motifs, d'intérêt pratique, pour la simplification des opérations de calcul. Nous, au contraire, nous avons calculé D et U pour tous les lemmes avant de choisir la limite inférieure. Du point de vue de l'élaboration avec l'ordinateur, leur procédé nous a semblé plus complexe. En fait le calcul de U et D pour tous les lemmes exige un temps d'élaboration un peu plus long, mais offre en échange l'avantage d'un choix plus réfléchi du seuil inférieur. En tenant compte de ces considérations nous avons procédé ainsi:

a) nous avons énuméré les lemmes dans l'ordre décroissant de F, de D et de U en les numérotant progressivement dans les trois ordres;

b) à l'exemple de JUILLAND, nous avons d'abord considéré la possibilité d'accueillir dans le LIF les premiers 5.000 lemmes dans l'ordre d'usage (U). Le lemme 5.000 a $U = 2{,}00$; les lemmes 4.945 à 5.001 ont cette même valeur;

c) à ce moment nous avons observé qu'en arrivant jusqu'au lemme 5.356, le dernier des 260 qui ont $U = 1{,}78$, on inclut tous les lemmes ayant $R > 3$.

Nous avons cependant fixé à $U = 1{,}78$ le seuil inférieur pour l'acceptation des lemmes dans le LIF. (Notez que 1,78 est la valeur de U pour $F = 3$ et $R = 3$).

En opérant ainsi nous avons ajouté aux 5.001 lemmes seulement 95 lemmes ayant $R < 3$ et 260 lemmes ayant $R = 3$.

Puisque entre le lemme 1 et le lemme 5.001 les lemmes ayant $R < 3$ sont 331, le LIF est constitué, en définitive, par 426 lemmes ayant $R < 3$ et par 4.930 ayant $R > 3$.

Une première confrontation avec le lexique de A. JUILLAND fournira des données précieuses sur les problèmes relatifs à la structure quantitative lexicale de la langue italienne; nous continuerons, ensuite, avec une comparaison des textes de notre bibliothèque du CNUCE, en remontant d'une part l'axe diachronique, et en nous déplaçant, d'autre part, sur l'axe synchronique vers d'autres styles et genres: prose philosophique, prose technique, textes populaires, textes « parlés » diversement enregistrés, etc.

Avec le LIF nous nous sommes proposés d'étudier la zone lexicale commune de quelques sous-ensembles de l'italien. Quelques uns de ceux-ci ont laissé apparaître une structure quantitative caractéristique, que nous nous proposons d'analyser maintenant, en les considérant à leur tour comme des univers lexicaux stratifiés en sous-ensembles. Nous commencerons probablement par la langue des journaux et par celle de la télévision.

3.3.2. *Les statistiques phonématiques de l'italien.*

« L'opera più estesa di statistica linguistica concernente il sistema fonologico italiano si deve a A. Zampolli (1960). Si basa sullo spoglio del dramma *Veglia d'Armi* di Diego Fabbri: il campione consta di 84.098 fonemi formati 37.117 sillabe e 18.970 parole » (Ž. MULJAČIĆ, 1968, p. 482)[1]. Les résultats de cette thèse n'ont pas été encore publiés. Ils seront intégrés à une recherche menée sur un corpus beaucoup plus vaste et varié.

N'importe quelle recherche statistique dans le secteur phonologique italien doit prendre conscience de deux problèmes de fond, de nature méthodologique: le choix de la prononciation pour transcrire les textes échantillons, et l'inventaire des unités phonématiques.

---

[1] Parmi les statistiques précédentes, je rappelle celles sur les lettres de VALERIO (1898) et celle de A. BOCCA e A. PELLEGRINI (1950). Cette dernière est faite sur des « sons » et non sur des lettres; cependant il est pratiquement impossible d'y identifier les phonèmes du système phonématique italien. De cette statistique les Auteurs ont tiré une liste de mots phonétiquement équilibrés, c'est-à-dire un ensemble de mots dans lequel la fréquence de chaque « son » s'approche de sa fréquence dans le corpus dépouillé; ces mots sont tous bisyllabiques (parce que la longueur moyenne trouvée par les Auteurs est 2,2) et paroxytons. Ces mots sont encore aujourd'hui employés pour l'examen médical de l'ouïe, surtout pour mesurer le degré d'adaptabilité sociale du patient. Il semble évident que cette liste pourrait être modifiée à l'aide de nos nouvelles données statistiques phonématiques, à l'aide des données sur les paires minimales, et à l'aide des données du LIF.

a) *Le choix de la prononciation.*

Il n'est pas possible de passer directement de la façon de parler individuelle à un seul type de prononciation nationale: l'italien a différentes prononciations régionales. Les études dans ce domaine viennent de commencer. Toutefois il est désormais vérifié que l'on ne pourrait pas parler d'un système phonématique italien unique, mais seulement de différents systèmes phonématiques régionaux. Les différences consistent dans les phonèmes utilisés et dans la distribution différente des phonèmes communs entre les mots. L'opposition, par exemple, entre les variétés de *e* et de *o* fermés et ouverts est connue dans de nombreuses formes parlées de l'italien. Cependant, on peut facilement énumérer une série de morphèmes qui sont prononcés dans une région avec le *e* ou bien le *o* fermés, et dans une autre région avec le *e* ou le *o* ouvert et viceversa. Les florentins prononcent *bène* (bien), *pième* (pied), *céntro* (centre), *accappatóio* (peignoir) et les vénitiens *béne*, *piéde*, *céntro*, *accappatòio*. (G. LEPSCHY, 1964).

Le choix d'un système régional de prononciation, conditionne donc les deux moments fondamentaux de l'enquête statistique. Il fait varier le nombre et l'inventaire des phonèmes (cas statistiques) dont on recherche les fréquences, et fait varier aussi la transcription phonématique, qui détermine la distribution des phonèmes du texte, c'est-à-dire la composition de l'échantillonage. Nous avons choisi le florentin savant, débarassé des caractéristiques et des nuances de la prononciation dialectale, parce qu'il offre les avantages suivants. La plus grande partie des études de phonématique ou de phonétique italienne prend comme base l'italien-florentin, si bien que la plupart des « tableaux » ou « listes » de phonèmes de l'italien représentent en réalité le système florentin. Deuxièmement, les dictionnaires attestent généralement la prononciation des lemmes selon l'usage florentin; en plus, celui-ci est sans doute le type d'italien le plus connu auprès des savants étrangers. Cela dit sans considérer que, sur la base d'une série de considérations historiques et de constations de fait, le florentin est proposé comme modèle orthoépique et présenté comme prononciation nationale. Citons à ce propos les propositions très claires du *Dictionnaire Encyclopédique Italien* (pp. II-IV).

Cela n'empêche pas que, selon une idée de G. DEVOTO, nous avons tracé un programme de travaux de statistique sur d'autres systèmes phonématiques régionaux ou dialectaux. En plus des textes enregistrés à dessein par le groupe de travail du *Centre pour la Phonétique du CNR*

de Padoue, dirigé par U. BORTOLINI, les dépouillements pourront profiter des matériaux recueillis dans le projet sur l'italien régional (3.3.3) et pour l'ALI.

b) *Inventaire des unités phonématiques.*

Le premier travail où l'on propose de manière explicite un inventaire des *phonèmes* italiens est sans doute l'article de G. PORRU, *Ammerkungen über die Phonologie des Italienisches* apparu en 1939 dans le huitième volume des *TCLP*.

Mais la discussion concernant le nombre des sons ou éléments de l'italien remonte à beaucoup plus loin.

Ž. MULJAČIĆ (1969, p. 410) la fait remonter « al lontano 1911 », c'est-à-dire au livre *Italiano. Fonetica, morfologia, testi*, de G. PANCONCELLI CALZIA, lequel parle des « 32 sons de la langue italienne, perçus par chaque italien instruit et doté d'une ouïe normale ».

M. SALTARELLI remonte même aux « *descrittivisti* » du XVI siècle (SALTARELLI, 1970, pp. 11 et suite). A ce propos, P. FIORELLI reconnaît la présence d'affinités remarquables entre l'analyse « phonémique » de A. CASTELLANI dans l'article *Fonotipi e fonemi in italiano* (Castellani, 1956) et les analyses du Toscan effectuées au XVI siècle par TRISSINO, TOLOMEI, GIAMBULLARI, RHOESUS, et, en particulier, par G. BARTOLI dans son *Degli Elementi del Parlar Toscano* (Firenze, 1584). Bartoli énumère 35 *éléments* qui sont en substance, selon FIORELLI, les 32 *phonotypes* ou allophones du CASTELLANI.

La discussion, stimulée par les données apportées par les recherches de phonétique historique du XIX siècle et surtout par les recherches expérimentales des phonétistes comme JOSSELIN (1900), PANCONCELLI CALZIA (1911), CAMILLI (1921), au cours de la première moitié de notre siècle, devient plus animée après la deuxième guerre mondiale, lorsque quelques travaux ont essayé de fixer, de manière plus ou moins explicite, le nombre des phonèmes italiens au moyen d'analyses menées selon les critères de la linguistique structurale.

A ce point-là, les divergences, entre les différentes inventaires proposés, apparaissent si importantes que l'on ressent l'exigence d'examiner ce fait et de l'expliquer.

On compare les différents systèmes proposés et l'on essaye de déterminer les causes des différences existant entre eux. Les premiers travaux effectués dans cette direction sont ceux de A. ZAMPOLLI (1960), et de G. LEPSCHI (1964), jusqu'à l'analyse très documentée et récente de Ž.

MULJAČIĆ dans le chapitre 8 de son oeuvre déjà citée *Fonologia generale e fonologia della lingua italiana* (1969).

La vieille controverse reste aujourd'hui plus que jamais ouverte à cause aussi de l'intervention récente dans la discussione des représentants de l'école générative, dont la position cependant est difficilement comparable, au moins en lignes théoriques, avec celles des études phonématiques précédentes.

Le nombre des phonèmes italiens oscille entre 28 et 53 dans les inventaires proposés jusqu'à nos jours. Cette constatation pose évidemment un problème complexe à quiconque s'apprête à effectuer, en dépouillant des textes, un relevé statistique des fréquences des phonèmes italiens.

A première vue, il semblerait inévitable que cette statistique doive être menée suivant un inventaire de phonèmes, et un seul, un choix préliminaire parmi tous les phonèmes proposés étant obligatoires.

Si cela était vrai, les résultats du relevé statistique seraient utilisés seulement par les savants qui sont d'accord avec ce choix et on devrait répéter le relevé autant de fois qu'il y a d'inventaires.

Mais un examen plus attentif de la situation nous a convaincu qu'il est possible d'obtenir des données statistiques utilisables pour tous les inventaires proposés. Un bref exemple nous permettra d'éclaircir cette affirmation.

L'on sait que, en ce qui concerne les consonnes soi-disant doubles ou géminées, « nella descrizione dell'italiano due punti di vista diametralmente opposti stanno combattendo già dalla fine degli anni trenta e la polemica scientifica non pare ancora esaurita ». (Ž. MULJAČIĆ, 1969, p. 427).[1]

Quelques phonologues estiment que ce qui distingue des mots comme *pala* (pelle) et *caro* (cher) des mots respectifs *palla* (balle) et *carro* (char) ce soit la présence, dans ces deux derniers mots, d'un phonème en plus. Les phonèmes de *pala* et *caro* seraient au nombre de 4, respectivement /p/, /a/, /l/, /a/ et /k/, /a/, /r/, /o/; les phonèmes de *palla* et *carro* seraient au nombre de 5, respectivement /p/, /a/, /l/, /l/, /a/ et /k/, /a/, /r/, /r/, /o/. Le même phonème /l/ apparaîtrait une fois dans *pala* et deux fois dans *palla*; le phonème /r/ apparaîtrait une fois dans *caro* et deux fois dans *carro*.

---

[1] Récemment quelques phonologues « génératifs » ont soutenu que dans la paire *pala* = *palla* serait distinctive l'opposition entre le /ā/ long de *pala* et le /á/ bref de *palla*, tandis que dans les deux mots on devrait reconnaître une occurence unique du phonème /l/.

D'autres phonologues soutiennent, au contraire, qu'il existe une classe spéciale de phonèmes appelés *longs* (ou encore *renforcés*, *intenses* et récemment *tendus*) qui diffèrent d'un même nombre de phonèmes appelés *brefs* (*non renforcés*, *non intenses* ou *relâchés*).

Ce qui distingue *pala* de *palla* et *caro* de *carro* ce serait précisément la présence dans *pala* et dans *caro* des phonèmes brefs respectivement /l/ et /r/, en opposition à la présence dans *palla* et *carro* des phonèmes longs correspondants /l// et /r//. Donc ces phonologues comptent seulement 4 phonèmes dans ces 4 mots.[1] On voit facilement que pour fournir des données statistiques utilisables par les deux points de vue opposés, il suffit d'introduire, dans l'inventaire des unités phoniques (UF) dont on décompte les fréquences, les UF ⟨l⟩ et ⟨l/⟩: la première apparait dans les mots du type *pala* (pelle), *calo* (diminution), *vale* (il vaut), etc.; la deuxième correspond au lien *ll* dans les mots du type *palla* (balle), *callo* (cal), *valle* (vallée).

En fait, il est évident que, pour les partisans de la deuxième interprètation, c'est-à-dire pour les phonologues qui soutiennent l'opposition consonne *brève*-consonne *longue*, la fréquence (*F*) de l'UF ⟨l⟩ coincidera avec la fréquence du phonème bref /l/, tandis que la fréquence de l'UF ⟨l/⟩ coincidera avec la fréquence du phonème long /l//.

Au contraire, les partisans de la première solution, c'est-à-dire les phonologues qui comptent dans le mot *palla* deux occurences du même

---

[1] Les conséquences du choix entre les deux solutions ne se limitent pas aux 15 consonnes en question, mais s'étendent à toutes les autres unités. En fait non seulement le nombre des phonèmes dans l'inventaire mais encore la longueur (*L*) en phonèmes du texte varie, longueur qui entre dans le calcul des fréquences en pour-cent.

Si nous appelons *M* le nombre des occurences globales des groupes soi-disant «consonnes doubles», et *N* le nombre global des phonèmes «non doubles», la longueur du texte est, si on adopte la solution monophonématique

$$L = M + N$$

tandis que si l'on adopte la solution biphonématique:

$$L = N + 2M$$

et donc la fréquence relative *f* d'un phonème *i* dont la fréquence absolue est F_i sera dans le premier cas:

$$fi = \frac{F_i}{M + N}$$

et dans le second cas au contraire:

$$fi = \frac{F_i}{N + 2M}$$

Notez que, d'après nos dépouillements, il résulte que:

$$M \simeq \frac{1}{10} N.$$

phonème /l/ obtiendront la fréquence de /l/ en ajoutant à la fréquence de l'UF ⟨l⟩ la fréquence multipliée par deux de l'UF ⟨l/⟩, selon la formule:

$$F\ /l/ = F\langle l\rangle + 2F\langle l/\rangle$$

c) *Le corpus échantillon.*

J'ai montré dans ma thèse qu'en subdivisant le texte échantillon *Veglia d'armi* en parts de 10.000 phonèmes chacune, les fréquences des UF dans chaque part étaient stables, comme l'a montré clairement l'application du test X^2.

D'autre part, il semble que d'autres données (la longueur moyenne des mots en syllabes, par exemple) dépendent au contraire du type de textes pris en considération.[1] Pour cette raison nous avons pensé à utiliser le procédé de stratifier le corpus qui se composera des textes recueillis pour le LIF, textes qui présentent l'avantage d'être déjà élaborés, lemmatisés et classifiés du point de vue morphologique, et quelques conversations de « locuteurs natifs » toscans, diversement choisis en considération du sexe, de l'âge et des conditions socio-culturelles.

Il s'agit, dans l'ensemble, de 600.000 mots, pour un total de 3.000.000 de phonèmes et de 1.300.000 syllabes environ.

d) *Le relevé des données.*

Le dépouillement des textes qui seront transcrits automatiquement au moyen du DMI concerne différentes unités (les traits distinctifs, les UF, les syllabes, les types syllabiques, quelques traits prosodiques, les structures syllabiques des mots etc.) et concerne aussi leur rapport syntagmatique [2] (leur position dans la structure du mot et de la syllabe, les séquences de 2,3 . . . *n* phonèmes, les séquences des types syllabiques, etc. . .). On détaillera la fréquence globale en sous-classes différentes, par exemple les différents sous-ensembles du corpus, pour évaluer la répartition, les catégories grammaticales, etc. En plus, on effectuera les calculs habituels des différents ordres d'entropie (H_0,H_1,H_2 . .).

Le travail dont nous prévoyons la publication en 1972 sera complété par des données correspondantes sur la structure phonématique du lexique. (cf. 3.2.3).

---

[1] Cf. par exemple N. TRUBETZKOY (1939) chap. 7.

[2] Voir par exemple le travail déjà cité de J. KRAMSKÝ (1964).

*

3.3.3. *Syntaxe et recherches socio-linguistiques.*

Le terrain de notre enquête est l'italien contemporain: on connaît assez bien l'état de déficit où se trouvent les études relatives à la description systématique de l'italien. L'analyse en a été faite plusieurs fois et il suffit ici de rappeler le fait. Avant de se lancer dans n'importe quel travail de type interprétatif, il est urgent de procéder au préalable à une reconnaissance systématique des types de signes existants en italien.

Puisque nous voulons que notre recueil s'étende aux domaines peu explorés des variétés régionales de l'italien nous avons étendu l'enquête à des textes d'italien populaire, à côté du matériel syntaxique reflétant l'usage écrit et standard de la langue italienne.[1]

Le but immédiat de cette recherche est double. Il s'agit d'une part d'intégrer les statistiques fondées sur les textes littéraires, ou de toute façon reflétant l'usage formel de l'italien, à des statistiques déduites de textes d'italien populaire. On tente d'autre part presque pour la pre-

---

[1] Le corpus sera ainsi constitué:

400.000 mots pris dans des textes échantillon d'italien populaire;
100.000 mots pris dans des devoirs d'enfants de « Scuola Media ».
500.000 mots pris dans des textes littéraires ou de toute façon reflétant l'emploi formel de l'italien.

La recherche est menée par T. De Mauro, G. Policarpi, G. Proverbio, A. Zampolli. Voir pour de plus amples détails T. De Mauro, G. Policarpi (1971). On veut porter finalement sous les projecteurs de la recherche scientifique des zones linguistiques importantes et encore dans l'ombre jusqu'à aujourd'hui: si, en fait, les études sur le diasystème italien en phonétique et en phonologie ne manquent pas, par contre les études sur les différences morphologiques et lexicales sont peu nombreuses et celles sur la syntaxe très rares.

Cette extension rattache notre recherche aux courants aujourd'hui prédominants dans les études sur la réalité sociolinguistique italienne, où le problème le plus important est l'intégration des immigrés. (cf. C. Grassi, 1964).

Au contact de la nouvelle société, le système ou les systèmes linguistiques de l'immigré entrent en crise face aux systèmes modèles du pays, et apparaît le besoin d'une organisation nouvelle.

Au moment où se produit l'impact avec la nouvelle réalité sociale et linguistique, l'immigré est profondément désorienté à cause du manque de modèles sociaux et linguistiques. Les conséquences, souvent négatives, de cette situation peuvent être considérées, par exemple, dans le milieu scolaire. L'étude de ces faits intéresse les sociologues pour des motifs évidents (voir par exemple F. Alberoni, G. Baglioni, 1965), mais fournissent aussi de précieuses informations sur la dynamique du processus dialecte – langue standard – variétés nationales, et des perspectives sur son résultat.

En même temps, on augmente la documentation, jusqu'alors maigre, sur la langue parlée, sur l'Italien populaire, et sur les variétés régionales.

Il faut rappeler, en particulier, les rapports sur les thèmes *Lingua e dialetto nella Scuola Italiana* et *Psicolinguistica, glottodidattica, linguistica contrastiva* présentés au IV Congrés International d'Etudes de la *Società di Linguistica Italiana*, Rome, juin 1970.

mière fois en Italie d'aborder la syntaxe de la période sous l'angle de la statistique. Après avoir rejoint ces deux buts on pourra affronter avec des instruments plus efficaces que ceux qui sont actuellement disponibles la rédaction d'une grammaire descriptive de l'usage effectif italien.

On sait que les recherches statistiques dans le domaine de la syntaxe sont beaucoup moins fréquentes que celles effectuées dans le domaine de la phonématique et du lexique.[1] Les raisons d'ordre scientifique, liées à l'état actuel et à l'histoire récente des études syntaxiques, soit générales soit se rapportant à des langues précises,[2] ne manquent pas. Mais les difficultés et les dépenses pour introduire dans le texte la description formalisée de sa structure syntaxique ont certainement aussi une grande importance.[3]

---

[1] Voir à ce sujet l'article de R. W. BAILEY, 1968.

[2] Voir par exemple l'article déjà cité de T. DE MAURO et G. POLICARPI, pp. 232-233. Ces dernières années, sous l'influence de la théorie de l'information de SHANNON, surtout, il y a eu quelques tentatives de description de la structure grammaticale d'une langue en termes statistiques (voir par exemple, CH. F. HOCKETT, 1967, p. 218). En partant des critiques de CHOMSKY sur le modèle marcovien qui englobe le concept de probabilité (N. CHOMSKY, 1957, pp. 21-22) et sur la relation entre la fréquence relative d'une phrase et sa grammaticalité (N. CHOMSKY, 1967, p. 34), l'école générativo-transformationnelle a nié toute fonction révélatrice à la statistique en grammaire et surtout en syntaxe: voir pour la discussion de cette position et pour la bibliographie, N. RUWET, 1968, pp. 39-40.

Toutefois, il me semble que si au lieu de parler de fréquence de phrases individuelles, on parle de fréquence de types syntaxiques, et on ne pose pas le problème des rapports entre fréquence et grammaticalité, mais on inventorie simplement les types existant et leur distribution en différents groupes de textes, en cherchant éventuellement des explications aux variations entre groupes et sous-groupes, les objections des générativo-transformationalistes devraient disparaître comme le montrent certaines études récentes, de cette école, sur le style. Ils partent de l'hypothèse que l'impression ressentie par le lecteur d'une oeuvre donnée vient en grande partie du type de procédés choisis par l'auteur, et qu'en particulier l'emploi des transformations (dans le sens du modèle générativo-transformationnel de Chomsky) est une partie très significative de ce que le lecteur perçoit comme style littéraire. La grammaire transformationnelle, en « formalisant » la notion de complexité des *sentences* superficielles en relation avec les *kernel-sentences* dont elles dérivent par transformation, permettrait de capturer un élément important constitutif de la notion de complexité stylistique et de le quantifier en mesurant l'emploi des appareils transformationnels dans un texte: par exemple, en comptant le nombre et les types de transformations appliquées dans chaque *sentence* et en confrontant les fréquences en pour-cent chez différents auteurs. Voir, par exemple, C. W. HAYES, (1966 et 1968) et R. OBMANN (1964).

[3] De ce point de vue, les difficultés sont plus grandes que dans les recherches phonématiques et lexicales. Dans ces deux secteurs, en effet, les unités du décompte statistique (phonèmes et mots) ont une représentation graphique directe. En syntaxe, au contraire, les unités du décompte sont les types syntaxiques: nous dirons que tandis que pour la phonétique et pour le lexique on opère sur les « suites primaires » des graphèmes qui

L'étude des fréquences des constituants de la proposition (par exemple: sujet, prédicat, objet direct, attributs, etc.), l'étude de leur succession, l'étude de la probabilité ou de la « prévisibilité » de l'occurrence d'un constituant à un certain point de la séquence linéaire, exigent en fait le *parsing* mot à mot du texte tout entier.

Ces recherches sont le plus souvent menées dans le contexte d'études sur le style, et prennent pour base les catégories des grammaires traditionnelles. Une grande partie des chercheurs est arrivée à la conclusion que les fréquences des principales catégories de constituants ne sont pas caractéristiques de chaque texte, mais plutôt d'un système linguistique tout entier.

Il semble au contraire que les fréquences des différents types de propositions varient sensiblement d'un texte à l'autre et d'un style à l'autre; c'est pourquoi elles semblent avoir un pouvoir de caractérisation et une valeur distinctive beaucoup plus élevée.[1] L'un des indices les plus efficaces semble être le rapport *s/p/d*, où *s* représente la fréquence des périodes simples, c'est-à-dire qui consistent en une seule proposition principale, tandis que *p* et *d* sont les fréquences respectives des propositions principales et des propositions dépendantes, dans les périodes composées de plusieurs propositions.

Pour notre recherche, il a fallu avant tout établir un tableau des types de classification syntaxique de l'italien. Ce tableau prévoit 90 classificateurs environ, répartis en 9 colonnes pour un maximum de 22 caractères.[2] On perfore le texte sur des fiches et on l'enregistre ensuite sur bande magnétique. L'ordinateur décompose le texte en mots

---

constituent le texte, pour la syntaxe, avant d'effectuer le décompte, il faut substituer à la suite primaire « une suite secondaire », constituée par l'enchaînement de classes grammaticales ou syntaxiques, et on doit, d'une façon ou d'une autre, en représenter la structure. A un niveau plus simple, cette analyse n'est pas toujours nécessaire: par exemple, pour l'étude des intervalles entre les répétitions des éléments linguistiques dans le texte. Pour l'enquête sur la longueur des *sentences*, mesurée suivant le nombre de mots la composant, il suffit de décider quels éléments graphiques du texte délimitent une phrase. La recherche des groupes binaires, ternaires, etc. de mots peut-être aussi faite sans une analyse syntaxique préalable. Et, en fait, ce sont là les types de recherches statistiques sur la syntaxe les plus répandues. (Voir, par exemple, les études de G. U. YULE, 1938, et de H. SPANG-HANSSEN, 1963). La lemmatisation suffit pour étudier la fréquence de co-occurence de catégories grammaticales déterminées, à moins qu'il ne faille un rapport syntaxique déterminé entre elles.

[1] Pour une troisième approche, qui est pour ainsi dire, l'intermédiaire entre les deux autres, voir J. C. CORBEIL (1971).

[2] Voir dans l'article déjà cité de T. DE MAURO et de G. POLICARPI la justification du choix des classificateurs qui appartiennent à la syntaxe dite traditionnelle.

et l'imprime, un mot par ligne, dans la première colonne d'un module, dessiné à propos, module dans lequel sont tracées les colonnes du tableau d'analyse, chacune comportant une en-tête. Les chercheurs effectuent sur ce modèle l'analyse syntaxique du texte: ils indiquent la division du texte en périodes, la division de la période en propositions, les rapports de coordination ou de subordination entre les propositions d'une période. Ils attribuent à chaque proposition des classificateurs syntaxiques qui la qualifient.[1]

Après avoir inséré cette analyse dans la bande-texte au moyen de fiches perforées l'ordinateur produit une nouvelle bande (bande-proposition) dans laquelle l'unité d'enregistrement est constituée par la proposition. En d'autres termes chaque *record* contient une proposition accompagnée par ses classificateurs et par un ensemble de renvois numériques qui situent la proposition dans la structure de la période dont elle fait partie; c'est-à-dire que ces renvois permettent de reconstruire la chaîne éventuelle de coordination et de subordination qui la relient aux autres propositions. On en déduit différentes statistiques syntaxiques: la fréquence de chaque classificateur syntaxique; les fréquences de « co-occurence » de sous-ensembles particuliers de classificateurs; la fréquence de différents types syntaxiques; la distribution des différentes longueurs des propositions, évaluées en mots; la distribution des différentes longueurs des périodes, évaluées en nombre de mots ou de propositions; les intervalles entre les répétitions des types syntaxiques; la fréquence des structures de la période en tant que séquence de types syntaxiques diversement enchaînés entre eux par des

---

[1] A l'intérieur de la période les propositions sont numérotées progressivement de telle sorte que chaque proposition est caractérisée d'une façon univoque par un numéro d'ordre.

Dans le modèle il y a deux colonnes réservées à l'indication des rapports de subordination et de coordination entre les propositions de la période.

Dans la première colonne, pour chaque proposition secondaire, on doit inscrire le numéro de la proposition dont elle dépend.

Dans la deuxième colonne sont, au contraire, reportées les données concernant le rapport de coordination que la proposition considérée établit avec celle du même type (principale ou secondaire) qui la suit dans la période: on inscrit le symbole indiquant le type de coordination et le numéro de la proposition coordonnée à celle analysée. A l'ordinateur arriveront ainsi les données relatives à la coordination entre les deux composantes des divers couples de propositions: l'ordinateur sera alors en mesure de reconstruire la chaîne éventuelle de coordinations qui relie les mêmes couples entre eux; aussi, ayant reçu les données relatives à chaque couple subordonnée-principale, il reconstruira soit la relation inverse qui n'est pas indiquée, soit la chaîne entière des dépendances dans la période.

relations de coordination et de subordination; les fréquences de sous-ensembles particuliers de ces enchaînements.

Nous obtiendrons même, en plus des concordances à exposant lexical, des concordances ayant pour exposants les classificateurs, ou leurs combinaisons, ou les types syntaxiques. Le contexte sera constitué dans chaque cas par la proposition toute entière à laquelle l'exposant se rapporte et, s'il le faut, aussi par les éventuelles propositions qui seraient en relation de coordination et de subordination avec cette proposition.

4. *Activité didactique.*

L'activité didactique s'exerce sous les formes suivantes:[1]

un cours de *Linguistique mathématique et computationnelle* auprès *de l'Istituto di Glottologia* de la Faculté des Lettres de l'Université de Pise.

un cours annuel de *programmes littéraires* auprès du CNUCE.

des bourses d'étude pour le perfectionnement en linguistique mathématique et « computationnelle ».

introduction des chercheurs des Instituts usagers à nos méthodes de linguistique « computationnelle ».

l'École d'été Internationale, *L'Elaboration électronique de données linguistiques et littéraires*, tous les deux ans. Nous décideront probablement d'alterner une Ecole à caractère général et propédeutique avec une Ecole spécialisée: pour la prochaine fois (1972) les arguments prévus sont:

A) *les grammaires formelles et les algorithmes pour le parsing*
B) *l'utilisation automatique des archives lexicographiques.*

---

[1] Cf. A. ZAMPOLLI (1969).

BIBLIOGRAPHIE

S. ABRAHAM, F. KIEFER, *An algorithmic definition of the morpheme*, in « Statistical Methods in Linguistics », IV (1965) pp. 4-9.

ACCADEMIA DELLA CRUSCA, *Concordanze degli Inni Sacri di A. Manzoni*, Firenze, 1967.

ACCADEMIA DELLA CRUSCA, *Novella del Grasso legnaiuolo. Testo. Frequenze. Concordanze*, Firenze, 1968.

F. ALBERONI, G. BAGLIONI, *L'integrazione dell'immigrato nella società industriale*, Bologna, 1965.

M. ALINEI, *Spogli elettronici dell'italiano delle origini e del duecento*, The Hague, 1968.

M. ALINEI, *Lessico e grammatica generativa*, in SOCIETÀ DI LINGUISTICA ITALIANA, *La grammatica, La Lessicologia*, Roma, 1969, pp. 73-96.

M. ALINEI, *Il tipo sintagmatico « quel matto di Giorgio »*, in SOCIETÀ DI LINGUISTICA ITALIANA, *Grammatica Trasformazionale Italiana*, Roma, 1971, pp. 1-12.

R. W. BAILEY, *Statistics and Style: A Historical Survey*, in L. DOLEŽEL, R. W. BAILEY (eds.) (1969), pp. 217-236.

M. BIERWISCH, *Über den teoretischen Status des Morphems*, in « Studia Grammatica », I (1962), pp. 51-89.

L. BLOOMFIELD, *Language*, New York, 1933.

E. BOCCA, A. PELLEGRINI, *Studio statistico sulla composizione fonetica della lingua italiana e sua applicazione pratica all'audiometria con la parola*, in « Archivio Italiano di Otologia, Rinologia e Laringologia » LVI (1950), suppl. n. 5, pp. 116-141.

U. BORTOLINI, C. TAGLIAVINI, A. ZAMPOLLI, *Lessico di frequenza della lingua italiana contemporanea*, IBM Italia, 1971.

R. P. BOTHA, *The function of the lexicon in transformational generative grammar*, The Hague, 1968.

R. BUSA, A. ZAMPOLLI, *Centre pour l'Automation de l'Analyse Linguistique (C.A.A.L.), Gallarate*, in *Les machines dans la linguistique*, Prague, 1968, pp. 25-34.

A. CAMILLI, *An Italian Phonetic Reader*, London, 1921.

J. B. CARROLL, *An Alternative to Juilland's Usage Coefficient for Lexical Frequences, RB-70-48*, Educational Testing Service, Princeton New Jersey, August 1970.

N. CHOMSKY, *Syntactic Structures*, The Hague, 1957.

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

N. CHOMSKY, *Topics in the Theory of Generative Grammar*, The Hague, 1966.

J. CORBEIL, *Les Structures syntaxiques du français moderne*, Paris, 1971.

T. DE MAURO, *Saggio di teoria formalizzata del noema lessicale (Appendice)*, in T. DE MAURO, *Introduzione alla semantica*, Bari, 1970, pp. 235-282.

T. DE MAURO, G. POLICARPI, *Ricerche sulla struttura del periodo italiano*, in SOCIETÀ ITALIANA DI LINGUISTICA, *L'Insegnamento dell'Italiano in Italia e all'Estero*, Roma, 1971, pp. 683-693.

R. J. DI PIETRO, *The Phonemic Status of Juncture in Italian*, in *Proceedings of the Fifth International Congress of Phonetic Sciences*, Basel-New York, 1965, pp. 261-263.

L. Doležel, R. W. Bailey (eds.), *Statistics and Style*, New York, 1969.

J. Dubois, *Étude sur la dérivation suffixale en français moderne et contemporain*, Paris, 1962.

J. Dubois, *Utilisation des statistiques lexicographiques pour l'étude structurale du lexique*, in *Statistique et analyse linguistique*, Paris, 1966, pp. 95-98.

A. Duro, A. Zampolli, *Analisi lessicali mediante elaboratori elettronici*, in *Atti del Convegno* sul tema: *L'Automazione elettronica e le sue implicazioni scientifiche, tecniche e sociali* (Accademia Nazionale dei Lincei, Roma, 1967), Roma, 1968, pp. 121-139.

Ch. J. Fillmore, *Lexical Entries for Verbs*, in « Working Papers in Linguistics », 2, The Ohio State University, (1968) pp. 23-24

W. N. Francis, G. M. Rubin, J. Svartivik, *A method of computer-produced graphical representation of dialectal variation in initial fricatives in Southern British English*, in « ICCL », 1969.

J. Friedman, *Applications of a Computer System for Transformational Grammar*, in « ICCL », 1969.

Dom J. Froger, *La critique des textes et son automatisation*, Paris, 1968.

J. Gardin, *A Typology of Computer Uses in Anthropology*, in D. Hymes (ed.), *The Use of Computers in Anthropology*, The Hague, 1965, pp. 103-117.

P. L. Garvin, *Computer Participation in linguistic Research*, « Language », xxxviii (1962) 4, pp. 385-389.

S. Golopenția-Eretescu, *La structure phonologique des monosyllabes roumains*, in « Cahiers de Linguistique théorique et appliquée », iii (1966), pp. 59-67.

J. Greenberg, *Essays in Linguistics*, Chicago, 1957.

A. J. Greimas, *Les Problèmes de la description mécanographique*, in « Cahiers de Lexicologie », 1 (1959), pp. 47-75.

P. Guiraud, *Bibliographie critique de la statistique linguistique*, Utrecht, 1954.

P. Guiraud, *Langage et communication. Le substrat informationnel de la sémantisation*, in « Bulletin de la Société de Linguistique de Paris », L (1954), pp. 119-133.

P. Giraud, *Les caractères statistiques du vocabulaire*, Paris, 1954.

P. Guiraud, *Problèmes et méthodes de la statistique linguistique*, Paris, 1960.

P. Guiraud, *Structures étymologiques du Lexique Français*, Paris, 1967.

D. G. Hays (ed.), *Readings in Automatic Language Processing*, New York, 1966.

D. G. Hays, *Introduction to Computational Linguistics*, New York, 1967.

D. G. Hays, *Computational Linguistics: Introduction*, in *Encyclopaedia of Linguistics, Information and Control*, Pergamon Press, 1969, pp. 49-51.

L. Heilmann, *Considerazioni statistico-matematiche e contenuto semantico*, in « Quaderni dell'Istituto di Glottologia » (Università di Bologna), vii (1962-1963), pp. 34-45.

C. W. Hayes, *A Study in Prose Styles: Edward Gibbon and Ernest Hemingway*, in L. Doležel, R. W. Bailey (eds.) (1969), pp. 80-94.

C. W. Hayes, *A Transformational-Generative Approach to Style: Samuel Johnson and Edward Gibbon* in « Language and Style », I (1968).

E. P. Hamp, *A Glossary of American Technical Linguistic Usage*, Utrecth-Antwerp, 1966.

G. Herdan, *Type-Token Mathematics*, The Hague, 1960.

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

L. Hjelmslev, *Pour une sémantique structurale*, in « Travaux du Cercle Linguistique de Copenhague », xii (1957).

C. F. Hochett, *A manual of Phonology*, Bloomington (Ind.), 1955.

C. F. Hockett, *The Quantification of*

Functional Load: A Linguistic Problem, Santa Monica (California), 1966.

C. F. HOCKETT, *Language, Mathematics, and Linguistics*, The Hague, 1967.

F. M. JOSSELIN, *Étude sur la phonétique italienne*, Paris, 1900.

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

A. JUILLAND, D. BRODIN, C. DAVIDOVITCH, *Frequency Dictionary of French Words*, The Hague, 1971.

A. JUILLAND, E. CHANG-RODRIGUEZ, *Frequency Dictionary of Spanish Words*, The Hague, 1964.

A. JUILLAND, P. M. H. EDWARDS, I. JUILLAND, *Frequency Dictionary of Rumanian Words*, The Hague, 1965.

W. J. HUTCHINS, *The Generation of syntactic Structures from a semantic Base*, Amsterdam-London, 1971.

L. GUILBERT, *Le dictionnaire du français contemporain*, in « Cahiers de Lexicologie », X (1967), 1, pp. 115-119.

M. KAY, *Standards for Encoding Data in a Natural Language*, in « Computers and the Humanities », I (1967), 5, pp. 170-177.

T. M. KNEASE, *An Italian Word List from Literary Sources*, Toronto, 1933.

J. KRÁMSKÝ, *A Quantitative Phonemic Analysis of Italian Mono-, Di-, and Trisyllabic Words*, in « Travaux Linguistiques de Prague », I (1964), pp. 129-143.

E. KUČERA, G. K. MONROE, *A comparative Quantitative Phonology of Russian, Czech and German*, New York, 1968.

S. M. LAMB, *The digital Computer as an Aid in Linguistics*, in « Language », XXXVII (1961), pp. 382-412.

S. M. LAMB, *Linguistic Data Processing*, in D. HYMES (ed.), *The Use of Computers in Anthropology*, The Hague, 1965, pp. 159-188.

G. C. LEPSCHY, *Prefazione alla edizione italiana dei Saggi Linguistici di Noam Chomsky*, in N. CHOMSKY, *Saggi Linguistici*, vol. I, 1969, pp. 9-17.

G. C. LEPSCHY, *Note sulla fonematica italiana*, in « Italia Dialettale », XXVII (1964), pp. 53-67.

J. LYONS, *Introduction to Theoretical Linguistics*, Cambridge, 1968.

A. MARTINET, *Éléments de linguistique générale*, Parigi, 1960.

B. MIGLIORINI, *La lingua del 900*, Firenze, 1941.

B. MIGLIORINI, *Der grundlegende Wortschatz des Italienischen*, Marburg, 1943.

B. MIGLIORINI, *Che cos'è un vocabolario?*, Firenze, 1951.

J. MISTRÍK, *Frekvencia slov v Slovenčine*, Bratislava, 1969.

C. A., MONTGOMERY, *The 1969 International Conference on Computational Linguistcs: A Progress Report*, in « Computers and the Humanities », IV (1970), 3, pp. 193-198.

R. MOREAU, *Au sujet de l'utilisation de la notion de fréquence en linguistique*, in « Cahiers de Lexicologie », III (1962), pp. 140-158.

R. MOREAU, *Intervention de M. R. Moreau*, in *Statistique et analyse linguistique*, Paris, 1966, pp. 125-132.

Ž. MULJAČIĆ, *Fonologia generale e fonologia della lingua italiana*, Bologna, 1969.

CH. MULLER, *Le mot, unité de texte et unité de lexique en statistique lexicologique*, in « Travaux de Linguistique et de Littérature », I (1963), pp. 155-173.

CH. MULLER, *Fréquence dispersion et usage: à propos des dictionnaires de fréquence*, in « Cahiers de lexicologie », II (1965), pp. 32-42.

CH. MULLER, *Initiation à la statistique linguistique*, Paris, 1968.

R. OHMANN, *Generative Grammars and the Concept of Literary Style*, in « Word », XX (1964), pp. 423-439.

M. PACAK, *Homographs: Classification and Identification*, in « Études de linguistique appliquée », V (1967), pp. 89-105.

G. PANCONCELLI CALZIA, *Italiano*, Leipzig-Berlin, 1911.

F., PAPP, *Mathematical Linguistics in the Soviet Union*, The Hague, 1966.

K. L. PIKE, *Language in Relation to a Unified Theory of the Structure of Human Behavior*, The Hague, 1967.

A. L. PROSDOCIMI, *Kutef persnimu ařepes arves: analisi interna e problemi redazionali nelle tavole iguvine*, in *Mille. I Dibattiti del Circolo Linguistico Fiorentino*, 1475-1970, Firenze, 1970, pp. 185-207.

D. W. REED, *A statistical Approach to quantitative Linguistic Analysis*, in « Word », v (1949), pp. 235-247.

J. REY-DEBOVE, *Limites des applications de la linguistique à la lexicographie*, in G. E. PERREN, J.L.M. TRIM (eds.), *Applications of Linguistics*, Cambridge, 1971, pp. 369-375.

J. REY-DEBOVE, *Le domaine du dictionnaire* in « Langages », 19 (1970), pp. 3-34.

I. I. REVZIN, *Les modèles linguistiques* (*Modeli jazyka*). Traduit et adapté par Y. Gentilhomme, Paris, 1968.

A. ROCERIC-ALEXANDRESCU, *Fonostatistica limbii române*, București, 1968.

A. ROSETTI, *Le Mot. Esquisse d'une Théorie générale*, Copenhague-Bucarest, 1947.

H. RUBENSTEIN, *Directions in Semantic Research*, in *Seminar on Computational Linguistics*, Bethesda, 1966, pp. 97-104.

N. RUWET, *La linguistique générale aujourd'hui*, in « Archives européennes de Sociologie », v (1964), pp. 277-310.

N. RUWET, *Introduction à la grammaire générative*, Paris, 1967.

M. SALTARELLI, *A Phonology of Italian in a Generative Grammar*, The Hague, 1970.

E. SAPIR, *Language*, New York, 1921.

N., SARAMANDU, *Sur le rendement fonctionnel des types de structures phonématiques en roumain*, in « Cahiers de linguistique théorique et appliquée », III (1966), pp. 147-160.

S. P. P. SCALFATI, *Notizie e studi a proposito della edizione delle pergamene pisane* (Secoli VIII-XII) in « Archivi e Cultura », IV (1970), 1-2, pp. 181-195.

R. W. SHUY, *An Automatic Retrieval Program for the Linguistic Atlas of the United States and Canada*, in P.L. GARVIN, B. SPOLSKY (eds.), *Computation in Linguistics*, Bloomington, 1966.

H. SPANG-HANSSEN, *The study of Gaps between Repetitions*, in MORRIS HALLE, et al. (eds.), *For Roman Jakobson*, The Hague, 1956, pp. 492-502.

H. SPANG-HANSSEN, *Sentence Lenght and Statistical Linguistics*, in *Structures and Quanta: Three Essays on Linguistcs Description*, New York, 1963, pp. 58-72.

H. SPANG-HANSSEN, *Fini et infini dans le vocabulaire*, in « Langages », 6 (1967), pp. 100-105.

J. STINDLOVA, *Le dictionnaire de la langue tchéque littéraire : enregistrement des données sur cartes mécanographiques*, in « Cahiers de Lexicologie », X (1967), pp. 103-113.

C. TAGLIAVINI, *Concordanze della « Divina Commedia »*, Pisa, 1965 (Ci si riferisce all'introduzione dell'edizione IBM Italia).

C. TAGLIAVINI, *Applicazione 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), Roma, 1968, pp. 111-118.

T. TORODOV, *Recherches Sémantiques*, in « Langages », I (1966), pp. 5-43.

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

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

J. VACHEK, J. DUBSKÝ, *Dictionnaire de linguistique de l'École de Prague*, Utrecht-Anvers, 1960.

R. L. WAGNER, *Les vocabulaires français*, Paris, 1967.

C. R. WOOD, *Dialectology by Computer*, in « ICCL », 1969.

G. U. YULE, *On sentence-Length as a*

Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Autorship in « Biometrica », XXX (1938), pp. 363-390.

A. ZAMPOLLI, *Studi di statistica linguistica eseguiti con impianti IBM*, (Tesi di Laurea dattiloscritta), Padova, 1960.

A. ZAMPOLLI, *Nota tecnica*, in « Raccolta Barbi di Canti Popolari Italiani. Esperimento di Elaborazione elettronica E 1/RB », Pisa, 1967, pp. II-XI.

A. ZAMPOLLI, *Intervento sul tema « Il dizionario italiano di macchina »*, in « Calcolo », V, suppl. n. 2 (1968), pp. 109-126.

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

A. ZAMPOLLI, *Due conversazioni sul panorama attuale della linguistica computazionale*, Pisa, 1969.

A. ZAMPOLLI, in *Cronaca. Notizie, spunti e appunti*, in « Archivio Glottologico », LV (1970), fasc. 1-2, pp. 272-279.

A. K. ZHOLKOVSKIJ, I. A. MEL'CHUK, *Sur la Synthèse Sémantique*, in « T. A. Informations », II (1970), pp. 1-85.