

Université de Paris 8
2, rue de la liberté
93200 Saint-Denis

Unité de formation et de recherche
Langage, informatique, technologie

Département Documentation

Contribution à une définition du domaine des
industries de la langue :
L'élucidation des statuts juridiques

~~Thèse~~ présentée en vue de l'obtention
d'un Doctorat en sciences de l'information
et de la communication

par

Christine Poitevin

Tél. 44.27.56.92
44.27.51.02
Fax 44.27.56.51

Table des matières

Introduction générale	page 5
1.- Les industries de la langue : enjeux, problématiques et produits	page 10
1.1.- les enjeux industriels	page 11
1.1.1.- De l'industrie de l'informatique aux industries de la langue	page 12
1.1.2.- Une automatisation encore partielle	page 17
1.1.3.- Les industries de la langue, prospective et innovation technologique	page 22
1.2.- Les enjeux culturels et stratégiques	page 27
1.3.- Introduction à une problématique du traitement automatique des langues naturelles	page 37
1.3.1.- Informatique et langue	page 37
1.3.2.- Langues naturelles et langues artificielles	page 42
1.3.3.- La linguistique et ses différentes branches	page 46
1.4.- Connaître la langue pour la traiter automatiquement	page 52
1.5.- Les industries de la langue : une classification fonctionnelle et une présentation des produits	page 64
1.5.1.- Les outils de description des langues	page 68
1.5.1.1.- Les dictionnaires informatisés	page 69
1.5.1.2.- Les dictionnaires électroniques	page 73
1.5.1.3.- Les corpus de références	page 82
1.5.2.- Le traitement de la langue : une proposition de hiérarchisation des fonctions	page 90
1.5.3.- Les produits issus des industries de la langue et leurs principales limites	page 105
1.5.3.1.- Ecouter et parler : les machines à dicter et les dialogues de commandes	page 109
1.5.3.2.- Lire et corriger : vérification et correction orthographique ou stylistique	page 121
1.5.3.3.- Lire, décrire et chercher : la constitution et l'interrogation de systèmes documentaires	page 130
1.5.3.4.- Lire et traduire : la traduction de textes	page 135
1.5.3.5.- Ecouter, parler, lire et corriger : l'assistance à l'enseignement	page 144

**Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques**

2.- La protection juridique des produits issus des industries de la langue	page 150
2.1.- Les conditions de la protection par le droit d'auteur, rappel des principes	page 153
2.1.1.- La création de forme, objet du droit d'auteur	page 154
2.1.1.1.- La non-protection des idées par le droit d'auteur	page 154
2.1.1.2.- L'étendue de la protection	page 159
2.1.2.- L'originalité de la forme	page 165
2.1.2.1.- L'originalité, condition nécessaire et suffisante	page 165
2.1.2.2.- La définition de l'originalité	page 169
2.2.- La protection du logiciel	page 172
2.2.1.- Eléments de définition technique d'un logiciel	page 173
2.2.2.- L'étendue de la protection et l'originalité du logiciel	page 176
2.3.- La protection des outils de description des langues	page 192
2.3.1.- Les dictionnaires à usage humain	page 192
2.3.1.1.- La "subjectivité" des ouvrages lexicographiques et terminologiques	page 192
2.3.1.2.- L'originalité des dictionnaires à usage humain	page 206
2.3.2.- Les corpus de références et les dictionnaires à usage d'automates	page 229
3.- Le statut juridique des produits issus des industries de la langue	page 239
3.1.- Contribution au débat sur la définition du logiciel	page 242
3.1.1.- L'élaboration des définitions du mot "logiciel"	page 242
3.1.1.1.- Les définitions dans les textes officiels	page 242
3.1.1.2.- Les définitions dans les dictionnaires	page 251
3.1.2.- Le choix des mots qui composent les définitions	page 255
3.1.2.1.- Une définition impressionniste	page 256
3.1.2.2.- Qui traite des données ?	page 259
3.1.3.- Qu'est-ce qu'un logiciel ?	page 260
3.1.3.1.- Codes de programmations et de commandes	page 262
3.1.3.2.- Des codes aux langages	page 268

**Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques**

3.1.4. - Langages naturels et langages artificiels, les critères de distinction	page 273
3.1.4.1.- Les oeuvres de langage et le critère d'apparence	page 277
3.1.4.2.- Les oeuvres de langage et le critère de finalité ..	page 279
3.2.- Une classification des oeuvres de langage	page 282
3.2.1.- les thésaurii	page 282
3.2.2.- Les requêtes à des S.G.B.D.....	page 284
3.2.3.- Logiciels et organigrammes	page 289
3.2.3.1.- Compilateurs, interpréteurs, langages de commande	page 289
3.2.3.2.- Les logiciels sont exprimés dans des langages artificiels	page 291
3.2.3.3.- Machines et oeuvres de langage	page 295
3.2.3.4.- Les manuels de programmation, les listages ..	page 296
3.2.3.5.- le <i>look and feel</i>	page 298
3.2.3.6.- Les organigrammes	page 302
3.2.3.7.- Les clones logiciels, critères de distinction ...	page 307
3.2.3.8.- Les logiciels inachevés et les logiciels qui ne fonctionnent pas	page 310
3.2.3.9.- La documentation	page 314
3.2.4.- Systèmes experts et bases de connaissances	page 318
3.2.4.1.- Proposition de définition d'un système expert	page 318
3.2.4.2.- Les systèmes experts sont-ils des logiciels ?	page 320
3.2.5.- Conclusions partielles	page 324
3.3.- Une classification des produits issus des industries de la langue .	page 327
3.3.1.- Les dictionnaires informatisés	page 327
3.3.2.- Les dictionnaires électroniques	page 328
3.3.3.- Les corpus de références	page 328
3.3.4.- Les oeuvres obtenues à la sortie d'un traitement par une machine	page 329
3.3.5.- Les logiciels	page 329
Conclusion	page 332
Bibliographie	page 341

INTRODUCTION GENERALE

Le traitement automatique du langage naturel, parlé par les humains, est devenu progressivement un champ d'applications industrielles. Désignées habituellement sous le nom d'industries de la langue, les activités de conception, de fabrication et de commercialisation de produits comme les vérificateurs orthographiques, les systèmes de synthèse de la parole ou de lecture optique de caractères écrits, ont connu un rapide développement depuis une trentaine d'années. En effet, ces activités se placent dans la cadre du profond remodelage qui est en cours dans l'ensemble du secteur tertiaire, sous l'influence, en particulier, des technologies de l'information.

Même si leur but ultime (faire mieux et plus vite que l'homme qu'elles sont destinées à remplacer) n'est atteint que dans peu d'applications, les machines à lire, à interroger des bases de données, ou même parfois à traduire, parviennent déjà à faciliter l'accès et à accélérer la circulation de l'information. Avant de se positionner dans une logique de remplacement pur et simple de l'humain (qui, néanmoins se profile déjà dans certaines applications), les industries de la langue proposent toutefois plus modestement d'améliorer sa productivité. C'est la mesure pragmatique de cette amélioration qui déclenche les investissements.

Mais les investissements nécessaires à la réalisation de produits du secteur des industries de la langue ne peuvent être consentis dans un environnement juridique qui n'apporterait pas un niveau de sécurité satisfaisant.

Aussi, le droit doit pouvoir apporter une réponse appropriée au besoin de protection du patrimoine intellectuel des personnes et des entreprises et remplir ainsi sa fonction régulatrice.

Les produits du secteur des industries de la langue sont des biens immatériels et peuvent, à ce titre, être considérés comme propriétés intellectuelles. Dans le droit positif, l'idée de propriété intellectuelle renvoie, entre autre, à celle de droit d'auteur et, plus particulièrement, aux Lois du 11 mars 1957 et du 3 juillet 1985. Cette thèse est donc centrée, pour sa partie juridique, sur l'étude de ces deux textes ainsi que des décisions jurisprudentielles qui les ont suivi.

Pour réaliser les fonctions de parler, écrire, traduire, les automates utilisent des fonds lexicographiques et terminologiques ainsi que diverses formes de descriptions des langues écrites et parlées. L'activité liée à la terminologie et à la lexicographie s'appuyant, pour une part importante, sur la "création de l'esprit", il est naturel qu'une protection juridique soit recherchée en premier lieu dans le régime des œuvres de l'esprit organisé par la loi du 11 mars 1957.

Mais ces ressources linguistiques sont, de plus en plus souvent, étroitement incorporées à des ensembles de logiciels et constituent des objets qui du point de vue du droit (d'auteur) relèvent de régimes juridiques différents, voire même contradictoires.

En effet, si la loi du 3 juillet 1985, en modifiant partiellement la loi du 11 mars 1957, a permis d'ajouter le logiciel à la liste des œuvres protégeables par celle-ci, elle a également soumis le logiciel à des dispositions particulières.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

La loi du 3 juillet 1985, en posant des règles propres pour les logiciels (copie non autorisée, droits dévolus à l'employeur et non à l'auteur-salarié du produit, durée de protection plus courte), est venue introduire un élément de distorsion au sein de la propriété littéraire et artistique et oblige à un effort de délimitation. Seul ce qui est proprement logiciel relèvera de la loi de 1957-1985 et tout ce qui n'est pas logiciel restera soumis à la loi de 1957 dans sa forme de droit commun.

"Il faut donc s'entendre sur le sens à attribuer au mot logiciel. Il le faut d'autant plus que les disparités auxquelles conduisent cette dualité de régime peuvent amener à des incohérences économiques quand ce qui est considéré par une entreprise comme un produit unitaire (un correcteur orthographique, un automate de traduction) relève de plusieurs régimes juridiques."¹

La réponse au problème du statut juridique ne peut être fournie qu'après s'être demandé si les produits issus des industries de la langue, considérés dans leur globalité et non à travers chacune de leurs composantes, peuvent être qualifiés de logiciels. En cas de réponse positive, la loi du 3 juillet 1985 leur sera appliquée. Dans le cas contraire, le statut légal de ces produits relèvera de la loi du 11 mars 1957.

La détermination du statut légal des produits issus du secteur des industries de la langue revêt une grande importance pratique car l'identification du ou des titulaires du droit de propriété sur ces produits, ainsi que la détermination de

1.- VIVANT Michel, LE STANC Christian, GUIBAL Michel et RAPP Lucien, *Lamy Droit de l'informatique*, n°851, Editions Lamy SA, Paris, 1992.

l'étendue de ces droits sont directement dépendants du statut envisagé.

Dans une première partie, nous proposerons une contribution à la définition du domaine des industries de la langue. Pour ce faire, nous prendrons d'abord la mesure des enjeux industriels, culturels et stratégiques de ce domaine. Nous constaterons, à cette occasion que ce secteur présente un certain nombre de caractéristiques qui rendent les quantifications et donc les prévisions très difficiles. Nous verrons également comment ce domaine repose principalement sur deux disciplines (l'informatique et la linguistique) dont l'une est dans un état d'avancement des connaissances qui est insuffisant pour aplanir les obstacles rencontrés lors de la conception de produits qui répondraient aux demandes des usagers. Cette difficulté actuelle à répondre aux besoins des usagers est, évidemment, source de conflits car les fournisseurs sont souvent amenés à promettre plus qu'ils ne peuvent. Mais nous ne traiterons pas dans cette thèse des conflits d'ordre contractuels entre clients et fournisseurs.

Afin de donner une base solide au raisonnement juridique, nous serons amenés à proposer, à la fin de cette partie, une typologie des produits issus des industries de la langue qui repose sur une analyse fonctionnelle de ces derniers.

Dans une seconde partie, nous étudierons les conditions de protection par le droit d'auteur des produits issus du secteur des industries de la langue. Se poser le problème du statut légal de ces produits, c'est admettre que ceux-ci doivent recevoir protection et que cette protection doit être exercée au titre du droit d'auteur. Mais ce postulat n'est qu'un point de départ. La protection par le droit d'auteur n'est, en effet, pas

acquise automatiquement et n'est qu'une potentialité. Il faut donc se demander si les produits issus du secteur des industries de la langue répondent aux conditions que le législateur impose pour accorder la protection. C'est pourquoi, après avoir développé les principes de protection (une création de forme originale), nous nous attacherons, reprenant la typologie définie en première partie, à identifier ce qui, dans chacun des objets, le rend susceptible d'être protégé.

Dans une troisième partie, nous apportons une contribution au débat sur la définition du logiciel. Ce concept, en effet, n'est volontairement pas défini en droit français, ce qui a pour effet de donner lieu à débat sur les textes et sur l'application des textes qui, pour leur part, utilisent ce terme non défini. Nous montrerons comment les méthodes habituellement utilisées par les juristes pour caractériser les logiciels (méthodes qui reposent essentiellement sur l'analyse des apparences) conduisent parfois à des résultats qui peuvent être discutés. Nous proposerons alors une méthode dont les fondements épistémologiques sont les mêmes que ceux de l'informatique, discipline dont nous avons envisagé l'objet principal comme étant la construction de langages artificiels en vue de la rédaction de modèles interprétables par des machines.

Pour terminer, nous appliquerons cette nouvelle méthode d'analyse aux différents produits issus des industries de la langue et nous constaterons alors son efficacité dans de nombreux cas qui, jusqu'alors, posaient problème.

Première partie

Les industries de la langue :
enjeux, problématiques et produits

1.- LES INDUSTRIES DE LA LANGUE : ENJEUX, PROBLEMATIQUES ET PRODUITS

Dans cette première partie, nous insisterons sur les problèmes spécifiques qui sont rencontrés lorsque l'on manipule les langues naturelles avec des ordinateurs. Nous décrirons les connaissances sur la langue qui doivent être formalisées pour les rendre exploitables par une machine. Nous terminerons par une classification des produits qui sont issus des industries de la langue, sur la base des fonctions qu'ils peuvent réaliser. Mais afin de mieux cerner le domaine des industries de la langue, nous nous proposons dès maintenant d'examiner les enjeux du développement de ce secteur. Si les enjeux sont d'abord économiques, ils sont aussi culturels et stratégiques.

1.1.- les enjeux industriels

A ses débuts, l'informatique avait principalement pour objet d'automatiser le calcul sur des nombres, qu'il s'agisse de balistique ou de comptabilité. Mais, progressivement, il est apparu que l'informatique pouvait également aider puissamment les hommes à gérer des stocks et des flux d'informations. On a ainsi pu assister au développement rapides de ces nouveaux objets que sont les banques de données bibliographiques, économiques, commerciales et juridiques.

Du mariage à trois entre l'industrie des télécommunications, l'industrie de l'informatique (avec ses deux composantes matérielle et logicielle) et une partie de l'industrie de la presse et de l'édition, sont nées ce que l'on a appelé les industries de l'information. L'apparition de ce terme a traduit le fait que l'information était devenue, en tant que telle, une matière première susceptible de traitements, d'affinages, de

regroupements et de commercialisations. Cela ne signifie pas pour autant que cette matière première aurait des caractéristiques identiques aux autres matières premières (en particulier, l'information a une caractéristique spécifique : quand on l'a diffusée, on la possède encore). Mais ce terme a permis de désigner simplement et simultanément les différentes facettes d'un concept complexe : l'aspect contenu et l'aspect contenant, l'aspect technologique et l'aspect commercial, l'aspect recherche et l'aspect applications.

Et surtout, ce terme a traduit le fait que les enjeux stratégiques et financiers liés aux industries de l'information devaient être considérés comme aussi importants que ceux qui sont liés aux industries de l'énergie, de l'agro-alimentaire ou de la sidérurgie.

1.1.1.- De l'industrie de l'informatique aux industries de la langue

Dès le début des années cinquante, lorsque l'informatique a commencé à se développer, l'idée est apparue que les ordinateurs allaient permettre de traduire des textes d'une langue à une autre. La guerre froide et la nécessité de lire les publications scientifiques en russe n'ont pas été étrangers au succès de cette idée dans tous les pays occidentaux. Initialement orienté vers la traduction automatique, le champ d'exploration s'est rapidement étendu à la vérification et à la correction orthographique, à l'indexation et au résumé de textes, etc.

Simultanément apparaissait l'impérieuse nécessité de concevoir des langages de programmation sans cesse plus évolués. A ces langages étaient nécessairement associés les

logiciels qui permettaient aux machines d'interpréter les textes des programmes afin de les traduire dans les langages de base des ordinateurs. C'est ainsi que les disciplines de l'informatique et de la linguistique ont été conduites à se rejoindre en une informatique linguistique, également appelée automatique linguistique ou encore linguistique computationnelle.

C'est par analogie avec le terme "industries de l'information" qu'est né en 1984, dans les services du Ministère français de la recherche, le terme "industries de la langue"¹.

Ce terme désigne "l'ensemble des activités qui visent à faire manipuler, interpréter ou générer par des machines le langage naturel écrit ou parlé par les humains"². Ces activités vont de la recherche au commerce en passant par le développement industriel de produits. Tout comme le terme "industries de l'information" recouvre (entre autres, car ce n'est pas le point le plus important du concept) l'activité qui consiste à faire de la recherche en sciences de l'information (bibliothéconomie, bibliométrie, économie de l'information...), le terme "industries de la langue" désigne (entre autres, car ce n'est également pas le point le plus important du concept) l'activité de recherche en informatique linguistique et en traitement de la parole.

Mais le secteur de la recherche n'est qu'une petite part du champ couvert par le terme "industries de la langue". Un autre sous-domaine important est celui du développement industriel

1.- DEGREMONT Jean-François, *L'émergence d'une industrie de la langue*, In : Brises, Paris, Editions du CDSH, n°4, Avril 1984, pages 5-6.

2.- CASSEN Bernard, *Les industries de la langue, un grand enjeu culturel, scientifique et technologique pour la France*, Rapport présenté à M. Hubert Curien, Ministre de la Recherche et de la Technologie - Paris, 7 Novembre 1985.

qui vise, partant des résultats de la recherche, à concevoir et réaliser des produits (matériels et logiciels) qui vont se vendre.

Il existe également un sous-domaine des activités de conseil en matière d'évaluation technique et commerciale et de conception de projets industriels incluant du traitement automatique de la langue naturelle. Pour désigner ces deux sous-domaines, le terme "ingénierie linguistique" traduit bien le fait qu'il est plutôt fait appel à des ingénieurs qu'à des chercheurs.

De nombreuses autres activités humaines sont également désignées par le terme "industries de la langue". Elles regroupent diverses professions qui sont basées sur l'exploitation des savoir-faire liés à la langue. Parmi celles-ci, il convient toutefois de distinguer les activités professionnelles qui reposent sur une maîtrise de la langue en tant que moyen d'expression (par exemple le journalisme) et celles qui portent sur la maîtrise de savoir-faire issus des sciences du langage (par exemple la traduction). Ces dernières recouvrent d'abord les métiers traditionnels du langage (de la traduction à la terminologie, en passant par la correction typographique) auxquels s'ajoutent un ensemble d'activités nouvelles : interrogation de banques de données, commande de robots, assistance à la manipulation de machines, etc.

Dans le terme "industries de la langue", le mot "industries" évoque la recherche d'une automatisation mise en œuvre dans le but d'accroître la productivité. En effet, les produits et les services qui sont offerts visent à introduire des traitements spécifiques de la langue dans des systèmes complexes en vue d'augmenter leur degré d'automatisation et, ainsi, de les

rendre moins dépendants des limites liées au travail humain : fatigue, erreurs... et coûts salariaux.

On peut également parler d'industries parce que les acteurs dans le domaine du traitement automatique des langues naturelles ont des tailles qui les placent clairement dans le champ de l'industrie et non dans le champ de l'artisanat. A cet égard l'exemple de la traduction est significatif et peut être généralisé : on est passé d'un stade artisanal, où des experts travaillent de façon isolée avec une productivité uniquement fonction du temps passé par le traducteur humain, à la mise en place d'organisations complexes utilisant de façon permanente les technologies avancées.

Enfin, le terme "industries" se justifie par le fait que les investissements nécessaires à la conception, à la fabrication et à la vente d'automates capables de manipuler le langage naturel sont souvent d'un niveau élevé et amènent à tenir les raisonnements caractéristiques des milieux industriels : équilibre financier entre investissement et marché potentiel, stratégies de commercialisation, emballage et maintenance des produits, recherche d'innovations effectivement commercialisables, si possible avec des effets de gammes, financement du développement par des marges sur les ventes et recherche d'une protection efficace contre des pillages éventuels.

Le fait que les composants linguistiques soient intégrés dans différents secteurs de l'industrie (introduction de modules de synthèse ou de reconnaissance de la parole dans les moyens de transport, les télécommunications, les machines-outils, de correcteurs orthographiques et de dictionnaires électroniques dans les systèmes bureautiques) n'interdit pas de parler d'industries de la langue en tant que domaine conceptuellement

indépendant : les industries de l'informatique sont, de ce point de vue, exactement dans la même situation que les industries de la langue. Il existe même un qualificatif pour désigner les industries ayant cette caractéristique de n'exister que par leur pénétration dans toutes les autres industries : ce sont des industries diffusantes.

Les industries de la langue "visent toutes un même objectif : contribuer à améliorer quantitativement et qualitativement les systèmes d'échange d'information, c'est-à-dire simplement la communication entre les hommes, entre les hommes et la machine, en attendant peut-être un jour la communication entre les machines."¹

Les industries de la langue apparaissent aujourd'hui comme l'un des domaines clés des technologies de l'information, à la fois parce qu'elles sont diffusantes (la plupart des autres domaines des technologies de l'information font ou feront appel à elles) et parce qu'elles sont incontournables (la plupart des autres domaines des technologies de l'information ne pourront se développer que conjointement à elles).

Ce phénomène est la conséquence d'un processus qui est analogue à l'explosion des communications téléphoniques à partir de 1850 : de 3 500 km en 1849, le réseau européen passe à 24 000 km en 1854, 65 000 km en 1859, 130 000 km en 1864, 7 000 000 km en 1913. En 1852, les six pays européens équipés (France, Allemagne, Autriche, Belgique, Royaume-Uni, Suisse) envoient 250 000 dépêches. En

1.- CHAUDIRON Stéphane, CONFLAND Daniel, *Le génie linguistique : contexte et développement*, Actes du Colloque "Génie linguistique 91", Versailles, France, 16 et 17 Janvier 1991.

1869 la France seule dépasse six millions de dépêches¹. Cette explosion des communications téléphoniques est gérée entièrement par des opérateurs humains : les célèbres standardistes du "22 à Asnières". Elle a continué sur un rythme soutenu pendant près de cinquante ans. Dès la fin de la seconde guerre mondiale, un calcul élémentaire montrait que pour offrir un téléphone à chaque foyer américain il faudrait bientôt que la moitié de la population féminine travaille dans un standard téléphonique ! Fort heureusement l'arrivée des centraux électromécaniques, puis des centraux numériques supprimait entièrement l'intervention des standardistes humains.

Dans le secteur tertiaire, des branches d'activité entières sont dépendantes de compétences linguistiques : le domaine de l'assurance repose, entre autre, sur un savoir-faire rédactionnel minimal, ainsi que sur la capacité à gérer un courrier important et varié. Contrairement aux banques, qui ont réussi à automatiser l'immense majorité des échanges épistolaires avec leurs clients, les assurances, aujourd'hui, gèrent manuellement leur courrier. Elles sont de ce point de vue littéralement asphyxiées.

1.1.2.- Une automatisation encore partielle

A ce jour, l'automatisation complète du traitement des langues naturelles est bien loin d'être acquise et, par conséquent, le goulot d'étranglement lié à l'indispensable présence d'opérateurs humains persiste. Les coûts de production qui résultent des difficultés à automatiser ont

1.- Numéro spécial de la revue INTERFERENCES. *Deux siècles de communication à distance*, C.N.A.M. et Association Internationale d'Histoire des Télécommunications et de l'Informatique, Paris 1985.

parfois des conséquences graves : en documentation scientifique et technique, seuls les Etats-Unis d'Amérique ont aujourd'hui les moyens de financer les équipes formées de centaines de documentalistes qui lisent, indexent et résument les considérables volumes de publications dans des domaines comme la chimie ou la médecine. On voit les conséquences pour les autres Etats, mis en situation de dépendance parce que contraints à acheter leur information de base et à pratiquer des stratégies de créneaux.

L'absence de théorie générale des langues naturelles, le faible niveau des connaissances sur les processus cognitifs humains ainsi que sur les mécanismes de la perception et de la production de la parole ne constituent cependant pas des obstacles à des réalisations partielles qui, sans automatiser totalement une fonction linguistique, permettent cependant de progresser dans la recherche.

La recherche ne peut actuellement apporter de solutions définitives à certains problèmes mais elle permet malgré tout des innovations : des automates spécifiques, aux performances linguistiques médiocres, peuvent rencontrer des marchés car ils améliorent, parfois nettement, la productivité des opérateurs humains, permettent d'offrir des services nouveaux ou même remplacent complètement un humain lorsque ses compétences linguistiques complètes ne sont pas requises.

Ainsi des vérificateurs orthographiques sont aujourd'hui disponibles sur tous les systèmes de traitement de textes. Ces vérificateurs détectent environ 40 % des fautes du clavier. Ce sont les fautes dites morphologiques (redoublement, omissions, inversions de lettres) à condition que la faute ne débouche pas sur un autre mot morphologiquement acceptable

(une chène de télévision, un cheik sans provision, les sports divers, un rapp or d'activité). Ces vérificateurs sont incontestablement très limités du point de vue de l'automatisation complète de la fonction de correction. Cependant ils améliorent non moins incontestablement la productivité de certaines catégories d'opérateurs humains, ce qui explique leur succès commercial.

On commence à voir apparaître la possibilité de piloter par la voix certains automates. Ainsi en est-il de la commande du microscope pour le micro-chirurgien ou de la commande des fréquences radio pour le pilote de chasse. Cette commande vocale ne fonctionne que sur des vocabulaires très limités et avec une syntaxe élémentaire. Elle n'offre en général pas une souplesse comparable à la commande manuelle. Elle permet cependant de libérer les mains pour d'autres fonctions plus importantes ou de pallier des déficiences fonctionnelles (hémiplegiques, non-voyants...).

L'interrogation de bases et de banques de données constitue actuellement le principal champ d'application des industries de la langue. Ici encore, l'automatisation complète est loin d'être acquise. Toutes les grandes banques de données textuelles fonctionnent avec des logiciels qui n'ont absolument rien à voir avec les compétences linguistiques d'un documentaliste : en effet, ces logiciels ne font pratiquement que comparer des chaînes de caractères.

En revanche, le documentaliste, même s'il n'est pas spécialiste du domaine d'interrogation, a des connaissances sur la forme comme sur le sens des mots utilisés par le demandeur (une recherche d'ouvrage traitant d'hôpital portera également sur la forme plurielle "*hôpitaux*", et inclura probablement ceux qui emploient les termes "*centre hospitalier*", "*centre de soins*"

intensifs", *"clinique"*, etc.), sur le but de ce dernier (un sociologue et un économiste, bien qu'interrogeant sur les mêmes objets, n'ont pas les mêmes buts), sur les circonstances de la demande (une recherche de documentation n'est pas la même pour un devoir d'étudiant, une thèse ou une étude industrielle), sur le "niveau culturel" du demandeur (on ne traite pas une demande d'un écolier comme celle d'un jeune étudiant ou d'un chercheur plus expérimenté).

En un mot, le documentaliste utilise de nombreuses compétences (relevant souvent du "sens commun", non connu et non perçu par un automate) pour gérer au mieux l'adéquation entre une demande et les multiples réponses possibles. Les ordinateurs, bien qu'ils répondent aux requêtes de façon souvent erronée, incomplète ou trop fournie, permettent cependant à une industrie de l'information (au chiffre d'affaires de plusieurs milliards de dollars par an) d'exister et de se développer. Leurs faibles performances sont partiellement contournées par l'imposition des contraintes linguistiques et documentaires à l'utilisateur et sont contrebalancées par des avantages certains : disponibilité, rapidité de mise à jour, exhaustivité.

En l'état actuel de nos connaissances, la recherche sur le traitement automatique des langues naturelles est, pour longtemps encore, un lent processus cumulatif. Si des avancées commerciales très rapides vont continuer à se produire (comme on l'a vu pour les vérificateurs orthographiques), elles apparaîtront dans le cadre d'un "grignotage" généralisé répondant petit à petit, de proche en proche, à des besoins précis (par opposition à l'idée d'une hypothétique machine à compétences linguistiques générales dont seraient dérivées, par apprentissage ou élagage, des machines spécialisées).

Soulignons, en particulier, qu'aucun expert en informatique linguistique pense que puissent arriver avant moins d'une vingtaine d'années des produits traitant de la langue sans contrainte (c'est-à-dire qui laissent l'utilisateur parler ou écrire comme il le souhaite, sans qu'il ait à s'occuper des capacités linguistiques de la machine), même si des maquettes visant ce but commencent à voir le jour dans certains laboratoires et même si cet objectif est inclus dans d'actuels grands programmes de recherche et développement (européens, japonais ou américains) à cinq ou dix ans.

1.1.3.- Les industries de la langue, prospective et innovation technologique

Dans les prochaines années, les marchés des produits et services du secteur des industries de la langue ne s'inscriront pas dans des logiques de satisfaction de besoins généraux mais dans des logiques de créneaux centrés sur des environnements particuliers. Ainsi, à défaut de savoir fabriquer un correcteur orthographique général, on fabrique des correcteurs spécialisés (sur les noms des rues, sur les marques commerciales...) ; à défaut de savoir fabriquer un système général de reconnaissance de la parole, on fabrique un composeur de numéros de téléphone ou une commande de robot à faible vocabulaire.

L'identification de ces créneaux et la conception des produits correspondants relèvent des habituels processus d'innovation et, en particulier, nécessitent qu'une culture technique en informatique et en linguistique se répande progressivement dans l'ensemble des lieux de production d'objets technologiques¹. Les prévisions sur les marchés futurs dépendent étroitement de l'idée que se font les prospectivistes de la vitesse des progrès de la technologie.

En suivant Jean-Jacques Salomon, nous partirons de l'idée que les progrès technologiques constituent des opportunités pour ceux qui savent les prévoir². Or la technique est relativement prévisible, car elle obéit à une logique de système³. La prévisibilité est certes relative : le catalogue des mauvaises

1.- Voir le développement sur ce thème dans "La révolution de l'intelligence, rapport sur l'état de la technique", numéro spécial 97 de la revue Science et Technologie, Paris, octobre 1983.

2.- SALOMON Jean-Jacques, *Le gaulois, le cow-boy et le samourai*, Etude CPE, ADITECH, Paris, 1985.

3.- GILLES Bertrand, *Histoire des techniques*, NRF-La Pléiade, Paris, 1978.

prévisions en matière d'innovation est constellé d'exemples célèbres. Dans le domaine des industries de la langue, nombreuses sont les études qui, au début des années soixante, prévoyaient l'arrivée de systèmes de traduction automatique de niveau industriel avant 1990. Ces erreurs ne signifient pas pour autant que toute prévision est impossible.

L'analyse des erreurs en matière de prévision montre qu'une grande part d'entre elles proviennent d'une sorte de myopie volontaire. Comme le dit le sociologue Pierre-Noël Denieul, "Dans les périodes de fort essor technologique, une bien faible place est accordée à une véritable prévision. Le présent y est trop dense, il est à lui seul l'avenir". Il y a une sorte de refus psychologique à admettre des conclusions qui seraient en contradiction avec la perception intime de l'analyste. Tout se passe comme si nous ne retenions des signaux qui nous parviennent de l'environnement que ceux qui sont à la fois conformes à nos connaissances et renforcent nos convictions, justifient nos (in-)actions, rencontrent notre imaginaire.

Si donc la prévision en matière d'innovation est difficile, on peut tenter de s'appuyer sur l'analyse d'autres mécanismes pour tenter d'identifier les lignes de force du développement technologique.

Quelques-uns de ces mécanismes ont été mis en évidence ces dernières décennies. Il est habituel aujourd'hui d'énoncer deux règles de base qui semblent fonctionner correctement¹ :

- les mutations technologiques sont des ruptures entraînant des déstabilisations. Le retour à l'équilibre, le rétablissement

¹ Voir, par exemple, GODET Michel, *Prospective et planification stratégique*, CPE-Economica, Paris, 1985.

de la cohérence s'effectuent, après chaque progrès technique important, lorsque celui-ci a eu le temps de diffuser dans l'ensemble du tissu industriel¹. Dans le domaine des industries de la langue, le phénomène de déséquilibre des compétences est particulièrement visible en ce moment : il y a, par exemple, un considérable décalage entre les technologies et les savoirs disponibles dans les entreprises qui fabriquent des logiciels de gestion de bases de données et celles qui fabriquent des dictionnaires à usage humain.

- la connaissance scientifique est utile mais n'est pas toujours à l'origine des innovations. Il est banal de rappeler que la machine à vapeur a précédé la thermodynamique, que le moteur électrique fonctionnait bien avant qu'on en connaisse les principes théoriques, que l'on savait fabriquer les alliages inoxydables avant de savoir pourquoi ils l'étaient. L'innovation se distingue clairement de la découverte scientifique et de l'invention : la découverte est la mise en évidence d'une loi naturelle (par exemple le pompage optique), l'invention est un principe nouveau de fonctionnement imaginé pour atteindre un objectif (par exemple le laser est l'invention utilisant la découverte du pompage optique). L'invention n'est pas nécessairement utile ni nécessairement réalisable de façon pratique. L'innovation technologique, en revanche, est l'application d'une idée nouvelle conduite jusqu'à sa réalisation industrielle, c'est la rencontre du champ du techniquement possible et celui du socio-économiquement possible.

Le secteur des industries de la langue peut s'observer à la lumière de cette définition de l'innovation technologique :

1.- *La Révolution de l'intelligence, rapport sur l'état de la technique*, numéro spécial 97 de la revue Science et Technologie, Paris, octobre 1983.

- il n'y a pas eu, à ce jour, de découverte permettant de mettre en évidence le fonctionnement de la langue naturelle, bien que de grands espoirs aient été conçus lorsqu'avaient été publiés les travaux de Zelig Harris, puis de Noam Chomsky sur les grammaires formelles. Un "Einstein de la langue" fournissant à brève échéance une "équation de la langue" n'existe toujours pas.

- il y a, par contre, eu de très nombreuses inventions. On peut citer de multiples cas de chercheurs ou d'entreprises ayant annoncé au cours des trente dernières années qu'ils avaient réussi à élaborer des mécanismes (souvent tenus secrets... pour cause de trop grande simplicité...) permettant d'indexer, de résumer, corriger voire traduire automatiquement des textes écrits. Ces inventions n'ont généralement pas réellement débouché : ce n'était que des inventions.

- il y a eu, pendant cette même période, beaucoup moins d'innovations, bien que l'on puisse cependant en décompter de fort intéressantes : les vérificateurs orthographiques disponibles sur les systèmes de traitement de texte, les synthétiseurs de parole à vocabulaires limités dans les jouets, les conjugueurs de verbes dans les systèmes d'enseignement assisté par ordinateur, les gestionnaires de dialogues en interrogation de bases de données structurées.

L'une des plus importantes conclusions du Rapport sur l'état de la technique est que l'évolution de l'innovation n'est pas linéaire.

Dans le domaine des industries de la langue, il apparaît clairement qu'il faut découpler, au moins partiellement, les analyses à propos de l'évolution de la recherche en traitement automatique des langues naturelles et les analyses portant sur

le développement et la commercialisation des applications industrielles. Le découplage n'est que partiel car c'est bien grâce aux succès commerciaux d'applications restreintes qu'est dégagée une partie des budgets pour des nouvelles recherches, qu'apparaissent de nouvelles pistes de travail, que sont expérimentés en vraie grandeur des prototypes. Cependant, si l'on peut envisager d'émettre des prévisions de marché dès qu'un produit commence à fonctionner, il reste très hasardeux de tenter l'exercice alors même que les recherches fondamentales n'ont pas encore abouti. On restera donc très prudent dans l'évaluation des marchés futurs comme des "prochains" développements techniques dans le domaine des industries de la langue, bien que nous pensions que ce secteur est appelé à jouer un rôle majeur dans l'évolution des industries de l'information.

Si, comme nous venons de le voir, les enjeux en matière de traitement automatique des langues naturelles sont d'abord industriels, ils présentent également des aspects culturels et stratégiques.

1.2.- Les enjeux culturels et stratégiques

La mondialisation de l'économie s'accélère depuis la fin de la seconde guerre mondiale et provoque la disparition des marchés locaux. Les matières premières puis les produits agro-alimentaires et les biens d'équipement sont aujourd'hui négociés sur une échelle planétaire. A cause des progrès en matière de transport et de distribution, cette planétarisation atteint même des produits inattendus : les tomates, les haricots verts et les fleurs, mais aussi les gadgets, les objets artisanaux et les vêtements d'occasion.

Parler aujourd'hui d'une extension et d'une interconnexion à l'échelle planétaire des réseaux de transmission de données et des médias (télévision en particulier) relève de l'évidence tant ce phénomène est puissant, bien que très récent.

En retour, ces réseaux d'information planétaires provoquent une interaction croissante entre les cultures et les peuples : l'émoi provoqué en quelques heures sur toute la Terre par les événements de la Place Tien an Men, de Timisoara, ou de la crise du Golfe en sont des marques récentes.

Par ailleurs, la mondialisation d'activités industrielles provoque l'apparition d'objets planétaires (le jean, la bouteille de Coca-cola, le 4X4 Toyota) auxquels la publicité se charge d'associer de nouveaux modèles de comportements.

Le brassage ne s'effectue pas seulement au niveau des marchés, des informations ou des objets. Il touche également ce que l'on désigne généralement sous le terme de culture. On a beaucoup parlé (et souvent en termes négatifs) de l'irruption des séries policières nord-américaines et des dessins animés japonais sur les écrans de télévision européens. On parle

moins d'autres phénomènes qui ont pourtant des ampleurs comparables :

- les musiques latino-américaines puis africaines sont devenues en moins de vingt ans une composante permanente des environnements sonores quotidiens en Europe.
- Des champs entiers de la science et de la technologie ont adopté une langue de travail commune, qui n'est pas toujours l'américain comme pourrait le faire croire l'observation du seul domaine de l'informatique. Ainsi, le français est largement utilisé dans les publications en physique nucléaire.

Le brassage concerne aussi les personnes physiques. Les déplacements de population ont pris une ampleur considérable : plus de soixante-quinze millions d'européens se sont installés sur le Nouveau Continent entre 1830 et 1920. Plus proches de nous, le développement des houillères du bassin rhénan dans les années 1830-1950 amena plus de cinq millions de migrants (polonais, italiens, turcs) en France et en Allemagne. Les guerres des années 1980 en Asie du Sud-est, en Afrique de l'Est, au Moyen-Orient ont provoqué des migrations qui, à chaque fois, se sont comptées en millions d'individus dans des délais très courts (généralement six mois à deux ans). Et que dire des immenses migrations estivales qui font venir chaque année pendant trois semaines en moyenne neuf millions de personnes en Grèce (pour neuf millions d'habitants), six millions en Thaïlande (pour cinquante millions d'habitants) et quarante trois millions de touristes en France. A côté de tels déplacements de populations, les invasions nordiques, "barbares" ou arabes qui pénétrèrent l'Europe par l'ouest, l'est et le sud du sixième au neuvième siècle de notre ère ne sont que petites excursions.

Ces brassages de grande ampleur ont des effets puissants sur les pratiques linguistiques : l'espagnol est pratiqué quotidiennement par environ 30 % des habitants des Etats-Unis ; des langues vernaculaires¹ sont littéralement atomisées par l'éparpillement des populations ; l'intégration d'Etats en des structures supra-nationales (l'URSS à partir de 1917, la CEE en ce moment) amène des tentatives de normalisations linguistiques à différents niveaux. L'analyse de ces processus est loin d'être simple. Ainsi, on pourrait imaginer que l'extension de la pratique d'une langue est liée à la puissance commerciale de ses pays d'origine. On pense évidemment à l'anglo-américain qui serait lié à l'impérialisme économique des Etats-Unis, succédant à l'impérialisme colonial du Royaume-Uni.

Cette hypothèse est partiellement exacte mais ne peut pas être généralisée. Elle est contredite par le cas de l'Allemagne et du Japon : ni l'allemand, ni le japonais ne sont en train de s'étendre alors que ces deux Etats sont actuellement aux plus hautes places en termes économiques et ont eu, dans le passé, des politiques coloniales agressives.

Des phénomènes d'ampleur non moins importante que les phénomènes économiques viennent complexifier le paysage. Ainsi, l'émergence des "identités nationales", la redécouverte

1.- On a l'habitude de distinguer les langues vernaculaires (langues maternelles) des langues véhiculaires, qui servent à la communication entre groupes de langues vernaculaires différentes. Le français est à la fois la langue vernaculaire d'une partie des habitants de la France (il y en a aussi dont la langue vernaculaire est l'espagnol, l'arabe, le basque), d'une partie des habitants du Canada, de la Suisse, de la Belgique, et la langue véhiculaire de l'aire dite "francophone". Le décompte des langues vernaculaires est rendu difficile par le fait que l'on ne sait pas très bien mesurer finement la distance entre les langues (et si même la notion de mesure a ici un sens) : il est clair que le français et le chinois sont des langues distinctes, mais les avis sont partagés lorsqu'il s'agit du français de France (région parisienne, pays de Loire, Provence ?), du français du Québec, des créoles des Caraïbes.

des histoires régionales, la montée des intégrismes religieux entraînent un renouveau des langues dont on aurait pu penser, pendant les trente glorieuses (1945-1975), qu'elles étaient mortes ou en voie de disparition¹.

Malgré la difficulté de procéder à des analyses fines, trois tendances culturelles et stratégiques lourdes semblent émerger de ces grands mouvements linguistiques :

- Après une longue période de diversification linguistique résultant de la propagation de l'espèce humaine à la surface de la Terre et des faibles communications entre les différents groupes, le nombre total de langues vernaculaires effectivement pratiquées sur Terre semble être en constante diminution depuis au moins un siècle. En France, on est ainsi passé de plus d'une dizaine de langues régionales à une seule langue nationale. Ceci ne veut pas dire que la disparition est totale ni que des sursauts ne se produisent pas, mais la tendance principale demeure.

- Moins d'une dizaine de langues sont puissamment renforcées par le fait qu'elles servent de langues véhiculaires dans des ensembles humains de dimensions planétaires. Ces langues sont, approximativement, celles qui sont admises comme langues de travail dans les institutions des Nations-Unies.

- Dans certaines de ces langues véhiculaires, on assiste à un fort développement des langages de spécialité. Contrairement aux "anciennes" langues, qui étaient attachées à des espaces géographiques ou politiques, ces langages de spécialité sont attachés à des professions. Ils sont peu utilisés en dehors de l'activité professionnelle, même si leur richesse lexicale est

1.- Hébreu, breton, basque.

tout à fait impressionnante : le vocabulaire technique du français ou de l'anglais est au moins dix fois plus important que le vocabulaire de la langue commune. Ne pas connaître et utiliser ce langage de spécialité est un facteur irrémédiable d'exclusion du champ professionnel. En effet, c'est en tout premier lieu par le fait que vous utilisez habilement cette langue de spécialité et par vos capacités d'innovations néologiques que vous démontrez à vos pairs votre maîtrise du domaine.

Les industries de la langue sont appelées à jouer un rôle amplificateur de grande importance pour ces trois phénomènes culturels et stratégiques.

Les trois exemples qui suivent le démontrent clairement :

- La rentabilité de la presse et de l'édition (qui constituent un soutien de première importance pour la pérennité et la diffusion d'une langue vernaculaire) est fonction directe de la disponibilité de machines à écrire. De nombreuses langues vernaculaires ont purement et simplement disparu depuis une centaine d'années parce qu'elles utilisaient des alphabets pour lesquels, faute de marché, on n'a pas fabriqué de machines à écrire, donc de composeuses typographiques mécaniques, donc de photocomposeuses électroniques. La presse écrite dans ces langues a perduré quelques années, les textes étant calligraphiés à la main. Puis elle a tout simplement fait faillite devant la concurrence des journaux composés avec des outils modernes dans les langues qui étaient "industrialisées" au niveau élémentaire de la machine à écrire. Ce mouvement de disparition est, évidemment, d'autant plus puissant que le nombre de fonctions linguistiques automatisées s'accroît dans la presse et l'édition.

- Pour de simples raisons de coût de transmission sur les réseaux de télex (qui ne fonctionnent aujourd'hui encore qu'en lettres capitales sur l'alphabet latin non accentué), il n'est pas concevable que les grandes agences de presse (Reuter...) diffusent leurs informations en plusieurs langues. La presse internationale, dans tous les pays du monde, est donc contrainte de puiser dans des sources qui sont pratiquement monolingues et de faire des traductions dans les langues nationales. A nouveau pour de simples raisons de coût, on cherchera à limiter le nombre de traductions et donc le nombre de langues de diffusion de ces informations. Et, par conséquent, les langues véhiculaires sont renforcées en permanence cependant que les langues vernaculaires sont affaiblies. Ce phénomène est très apparent dans les instances supra-nationales : malgré l'obligation légale de traduire tous les textes produits par la Commission des Communautés Européennes dans toutes les langues officielles des Etats européens, le volume de traduction ne peut suivre le volume de production (d'autant que le nombre de langues augmente avec l'arrivée de nouveaux Etats membres). Les langues pour lesquelles des systèmes d'aide à la traduction et à la mise en page existent sont, de fait, plus utilisées que les autres.

- Dans les domaines de l'aéronautique et de l'espace, les impératifs de sécurité d'exploitation, de maintenance, de coordination industrielle, de normalisation technique ont primé sur les faits linguistiques nationaux pour imposer une langue de spécialité (plus ou moins dérivée de l'anglais, mais incompréhensible par des anglophones non spécialistes de ces domaines). Ce langage de spécialité a fait l'objet d'efforts linguistiques particuliers (élaboration de dictionnaires sous forme électronique, de logiciels d'aide et de contrôle des rédactions techniques, de banques de données bibliographiques, textuelles et factuelles), dont les coûts importants

n'autorisent pas la duplication en d'autres langages (par ailleurs sans intérêt, puisque "tout le monde" dans la profession pratique le langage de spécialité). L'automatisation, même partielle, de traitements linguistiques dans ce langage réduit donc sans cesse la possibilité pour d'autres langues de développer dans ce même domaine un langage de spécialité.

Cet exemple de l'aéronautique est généralisable à l'ensemble des sciences et des techniques, bien que la situation soit loin d'être partout aussi tranchée : l'activité de recherche ne peut seulement être considérée comme s'exprimant dans le langage de spécialité. Elle est également très dépendante de la culture dans laquelle baigne quotidiennement le chercheur, à cause de l'étroite relation qui existe entre l'utilisation d'une langue et les modes de raisonnement scientifique¹. Cependant, les connaissances de la science et de la technologie n'existent que si elles sont exprimées, et cette expression passe par l'utilisation de langages de spécialité enchâssés dans des langues naturelles déjà importantes parce que véhiculaires. Aucune discipline - même les mathématiques - n'a pu se contenter d'un formalisme artificiel et s'affranchir complètement de l'usage d'une langue naturelle. Ainsi, la survie de toute langue en tant que langue scientifique et technique dépend-elle étroitement d'une part des possibilités de lui faire subir des traitements automatiques et, d'autre part, de la possibilité pour cette langue de s'imposer dans des champs scientifiques et techniques aussi nombreux que possibles.

1.- *Quelles langues pour la science ?* sous la direction de Bernard CASSEN - Paris, Ed. la Découverte, 1990, 263 pages.

Les industries de la langue semblent donc avoir un effet massivement normalisateur et réducteur de la diversité linguistique en favorisant l'accroissement des écarts entre langues majoritaires et langues minoritaires.

Il est cependant possible d'adopter un autre point de vue qui tempérera la première impression. On peut considérer que la diversité linguistique à la surface de la Terre, à l'échelle d'un continent, comme l'Europe, ou dans un pays unique, comme la Confédération helvétique ou l'Inde, constitue un handicap. On peut également en prendre acte comme d'un simple constat. On peut encore, malgré les tendances lourdes énoncées ci-avant, la promouvoir comme une ressource unique en l'exploitant avec discernement. On considère alors que c'est de la diversité linguistique et culturelle que naît la pluralité des points de vues sur le monde. Et le respect de cette diversité favorise une créativité accrue dans tous les domaines, et aussi permet l'apprentissage si nécessaire de l'écoute de la différence.

De ce point de vue, le développement des industries de la langue constitue alors un défi majeur pour l'humanité. Une automatisation accrue des traitements linguistiques dans le plus grand nombre de langues permettrait, en effet, de concilier l'accroissement des besoins de communication avec le maintien de la pratique des diverses langues. Et l'on peut imaginer que se pose moins aux pratiquants des langues minoritaires le dilemme de l'exclusion. Ceci aurait pour effet de revitaliser les langues minoritaires ou, en tout cas, de limiter la baisse de leur pratique. Ceci aurait surtout pour conséquence de remplacer par une logique pluriculturelle et plurilingue la pratique ancienne qui fait de la langue imposée un outil de domination.

Si le premier pilier de ce point de vue est le soutien à une vision pluriculturelle du monde (avec, en exemple évidemment caricatural, des conséquences sur les politiques étrangères des Etats : l'Europe de la diversité qui accepterait la différence aussi à l'extérieur de ses frontières contre le creuset uniformisant des Etats-Unis d'Amérique qui tenterait d'imposer son mode de vie à toute la planète), le second est un principe d'égalité de droit entre les Etats, les régions, les aires linguistiques, indépendamment de leurs puissances économiques ou politiques. C'est ce principe qui oblige la Commission des Communautés Européennes à traduire ses textes officiels en grec ou en danois comme en allemand ou en anglais. Bien entendu, ce principe se heurte aux contraintes purement économiques. Les efforts de recherche et développement sont du même ordre de grandeur quelle que soit la langue. Cependant, un vérificateur orthographique pour traitement de texte en anglais a un marché potentiel de plus de quatre cents-millions de locuteurs, contre environ neuf millions pour le grec et cinq millions pour le danois.

La pure logique commerciale étant à l'évidence contradictoire avec le choix politique, on comprend pourquoi il est alors nécessaire de raisonner le développement des industries de la langue dans des logiques stratégiques (peu importe le coût, puisque c'est vital) ou de service public (peu importe le coût, puisque c'est un droit).

Ces deux piliers (pluriculturalisme et principe d'égalité de droit entre les langues) constituent la base de la stratégie du Conseil de l'Europe et de la Commission des Communautés Européennes en matière d'industries de la langue.

Par ailleurs, de nombreux Etats ont compris que s'ils n'industrialisaient pas eux-mêmes leur langue, ils prenaient le

risque de voir d'autres Etats le faire à leur place (dans le cas des langues majoritaires), ou bien ils prenaient le risque de se voir progressivement colonisés par des langues déjà industrialisées (dans le cas des langues minoritaires). Ainsi, on a vu naître de nombreux programmes nationaux dans le domaine des industries de la langue : plan "filière électronique" en France et en Allemagne, programmes Alvey en Grande-Bretagne, DARPA aux Etats-Unis, cinquième génération au Japon...

1.3.- Introduction à une problématique du traitement automatique des langues naturelles

Plusieurs disciplines sont concernées par le traitement des phénomènes généraux qui caractérisent les langues. Elles soulèvent, selon des problématiques distinctes, diverses questions fondamentales et utilisent, pour les résoudre, des approches complémentaires. Une compréhension complète de ces phénomènes implique la prise en considération de tous ces points de vue différents dans un cadre unifié qui permette leur coopération. En effet, on conçoit facilement que, dans le projet de construction de machines parlantes, on aura impérativement besoin des résultats fournis par les psychologues (par exemple, pour concevoir l'ergonomie des dialogues : le "comportement" des machines comme suite à l'analyse par les machines du comportement des opérateurs humains), par les sociologues et les ethnologues du langage (il faut être capable d'analyser les comportements linguistiques de différentes "tribus"), et bien entendu par les linguistes (phonéticiens, phonologues, syntacticiens, sémanticiens), les électroniciens et les informaticiens.

1.3.1.- Informatique et langue

Si la dépendance de l'informatique par rapport aux autres disciplines est évidente, réciproquement, une partie de ces disciplines, qui étudient la façon dont l'homme et les sociétés humaines acquièrent, utilisent et génèrent des connaissances sur le monde, a besoin de l'informatique tantôt à titre d'outil de traitement des corpus étudiés (l'informatique permet d'aider à "classer" et à "mesurer" les langues), et tantôt comme support de simulations et vérifications d'hypothèses et de modèles. Les utilisations de l'informatique dans les sciences humaines et sociales sont aujourd'hui considérables.

La convergence entre ces disciplines sur un objectif commun (l'étude et la modélisation des mécanismes de la connaissance, sous toutes ses formes exprimées chez l'homme : apprentissage, compréhension, langage, perception, planification...) a amené à les désigner globalement sous le terme de sciences cognitives. Le projet commun à l'informatique et à la linguistique à propos du langage naturel est, de ce point de vue, très représentatif de ce que sont les sciences cognitives.

Les rapports qui peuvent exister entre informatique et langue sont de plusieurs ordres.

Tout d'abord, on peut considérer que l'informatique fournit des outils permettant une progression des connaissances sur la langue. Constructions de corpus (des banques de sons, des banques de textes), comptages et dénombrements, ordonnancements et classification, analyses statistiques diverses sont réalisés grâce aux ordinateurs. Ces opérations d'inventaire constituent la base de la démarche scientifique.

Parce que le chercheur dispose de machines sans cesse plus puissantes, il peut affiner ou remettre en cause plus facilement des catégorisations que la tradition lui avait imposé. Et l'on voit ici apparaître immédiatement le fait que ces activités constituent en elles-mêmes une source considérable d'idées. Ainsi on a l'habitude, en grammaire, de distinguer les noms communs et les adjectifs. Lorsque des dénombrements exhaustifs sont tentés à l'aide d'ordinateurs, la distinction entre ces deux classes de mots n'est pas toujours aisée, ce qui peut conduire à proposer de classer les mots suivant d'autres critères. Dans les activités d'inventaire et de classification, l'informatique joue principalement un rôle d'assistance.

Mais l'informatique peut également aider à comprendre le fonctionnement du langage grâce à des simulations de divers ordres. En effet, si la première étape de la démarche scientifique est l'inventaire et la classification, la seconde consiste à imaginer des hypothèses sur le fonctionnement de la langue (ou sur le fonctionnement des humains en train de parler), à formaliser ces hypothèses et à les inclure dans des automates et des simulateurs. En faisant fonctionner l'automate, on peut valider, réfuter ou modifier l'hypothèse initiale.

A ce jour, aucun ordinateur n'a pu être programmé de telle façon qu'il ait un comportement linguistique (dans une des langues naturelles) aussi élaboré que celui d'un humain. Dans des contextes restreints, des résultats satisfaisants ont été obtenus. Ces résultats paraissent faiblement significatifs dans un environnement ayant toute la complexité de la réalité. Malgré ces difficultés nous ne pouvons abandonner l'usage des ordinateurs pour traiter les langues. En effet, il y a une relation extrêmement étroite entre une théorie, un modèle de la langue et la machine, l'automate, qui matérialise cette théorie et permet de la valider. Et quelles que soient les contraintes liées à son usage, il se trouve qu'à ce jour l'ordinateur est le support le mieux approprié pour recevoir l'implantation de modèles de la langue et de modèles de l'être humain en train de manipuler la langue.

Ce point peut évoluer, par exemple si l'homme vérifie l'hypothèse que certains gènes contrôlent des fonctions linguistiques chez des êtres vivants. En suivant ces voies, il serait éventuellement possible d'utiliser d'autres supports que l'ordinateur pour construire des machines à traiter les langues humaines. Mais l'état des recherches ne le permet pas encore.

Un autre argument en faveur de recherches sur l'utilisation d'ordinateurs pour traiter le langage naturel est que c'est avec des systèmes à base d'ordinateurs que l'homme a aujourd'hui besoin de communiquer, depuis la machine-outil jusqu'à la banque de données.

Pour l'instant, nous continuons donc notre travail avec les outils dont nous disposons. D'une part parce que, rétrospectivement comme prospectivement, la voie semble féconde puisqu'elle continue à engendrer, à un rythme régulier, des machines sans cesse plus performantes. D'autre part, parce que, au-delà des débats sur une hypothétique stérilité future de cette voie (en particulier lorsque la complexité du monde réel est abordée), nous ne disposons pas de machines autres que les ordinateurs pour implanter des modèles artificiels de la langue naturelle qui permettent la simulation. On voit que la fécondité potentielle de cette voie prend deux aspects également importants :

- pour les informaticiens (concepteurs de machines qui sont mises à la disposition des autres sciences), aborder le traitement du langage naturel est source d'innovation dans la façon de concevoir machines et logiciels.

- pour les sciences du langage, l'utilisation des ordinateurs, lorsqu'elle est possible, permet d'accomplir le pas essentiel de la vérification des hypothèses formalisatrices et d'améliorer la productivité des chercheurs afin de progresser dans la connaissance des langues.

Dans le cadre des essais d'implantations d'automates traitant la langue naturelle sur des ordinateurs, deux principaux axes de travail sont possibles. Le premier considère la langue comme un système de représentation de connaissances, de

façon pratiquement indépendante du fait qu'elle ait été produite par des êtres humains. On s'efforcera, alors, d'analyser et de décrire l'organisation des signes (la morphologie, la syntaxe) et l'organisation des signifiés (la sémantique) de segments de textes et de paroles qui n'ont pas été principalement produits en situation de dialogue.

Le deuxième axe considère la langue en tant que vecteur essentiel de la communication homme-machine ; c'est l'étude des mécanismes de dialogue. Bien entendu, ces deux axes de travail se fertilisent mutuellement et s'entrelacent sans cesse.

L'informatique présente, à notre avis, une caractéristique qui est tout à fait unique : cette discipline a, avant tout, pour but de faire fonctionner des ordinateurs qui doivent être programmés. De ce point de vue, l'interlocuteur principal de l'informaticien en train de présenter ses connaissances, c'est son ordinateur, alors que l'interlocuteur principal des chercheurs dans les autres disciplines est le collège de ses pairs. Autrement dit, les informaticiens doivent expliciter, dans des langages interprétables par les ordinateurs, les tâches à réaliser, les façons d'appréhender le monde et de réaliser ces tâches, les objets du monde. Et ces langages, qui servent à expliciter pour la machine le monde qui l'entoure et les actions à entreprendre, ne sont pas des langues naturelles mais des langages artificiels, entièrement inventés par l'homme.

Aucune autre discipline n'a cette impérative obligation de tout décrire dans des langages artificiels. Même les mathématiques, la physique, la chimie, qui ont largement développé des langages artificiels pour décrire le monde et ses transformations, font majoritairement appel au langage

naturel¹. Quant aux sciences de l'homme et de la société, elles expriment leurs résultats quasi uniquement en utilisant le langage naturel. Il n'existe pas aujourd'hui de formalisme artificiel "universel", c'est-à-dire qui pourrait être indifféremment utilisé par toutes les disciplines. La faisabilité et l'utilité même de ces langages artificiels est, dans toutes les sciences de l'homme et de la société, un sujet de débat continu. Seules quelques branches des sciences de l'homme et de la société ont mis au point (la linguistique formelle) ou emprunté à d'autres disciplines (l'économie) des langages artificiels pour tenter de décrire leur objet de recherche.

Pour ce qui est de la linguistique, d'une part les langages artificiels qui ont été élaborés ne permettent pas aujourd'hui de décrire l'ensemble du champ de recherche (la sémantique et la pragmatique ne sont pas formalisées aujourd'hui) et, d'autre part, les langages artificiels élaborés ne sont pas stabilisés, c'est-à-dire ne rencontrent pas l'accord de la majorité des linguistes et ne jouent pas le rôle de communication "objective" des informations entre les chercheurs, comme le joue le langage que se sont construits les mathématiciens.

1.3.2.- Langues naturelles et langues artificielles

Qu'est-ce qui distingue aujourd'hui les langages naturels des langages artificiels ?². Les savoirs de la science et de la technologie ont cette particularité que, par définition, ils n'existent que s'ils sont verbalisés. Ils se distinguent ainsi de

1.- *Penser les mathématiques*, sous la direction de Françoise GUENARD et Gérard LELIEVRE - Paris, Ed. du Seuil, 1982.

2.- Ce point est traité également, d'un point de vue plus technique, dans la troisième partie de cette thèse.

nombreux autres savoirs, tels que le savoir-faire de l'artisan, le savoir du compositeur, du peintre, du danseur, le savoir de la vie quotidienne, qui peuvent être enseignés, utilisés et reconnus sans être nécessairement passés par une verbalisation.

Cette particularité découle du principe que la connaissance dite scientifique doit, pour être reconnue, être indépendante de celui qui la possède et l'utilise. C'est cette indépendance qui est évaluée lorsqu'une découverte est annoncée (la mémoire de l'eau, la fusion froide mais aussi la supraconductivité à haute température). Et l'évaluation, la reconnaissance, ne peuvent être réalisés que si la connaissance est extériorisée, par celui qui l'a découverte, et exprimée dans un texte, des formules ou des images duplicables et examinables par tout un chacun. Pour la science, il n'y a donc pas de connaissance sans expression de cette connaissance. Or l'expression passe par l'usage de langues naturelles, que ce soit la langue maternelle ou des langues véhiculaires, accompagné, dans certains domaines, de l'usage de langages artificiels.

Les langages artificiels (utilisés en mathématiques, en chimie, en informatique...) permettent de décrire des segments de la connaissance du monde qu'ont les chercheurs de certaines disciplines. Pourquoi donc avoir créé ces formalismes artificiels pour décrire les connaissances scientifiques, alors que le langage naturel des humains est d'une souplesse et d'une puissance apparemment immenses ? C'est parce que cette souplesse et cette puissance d'expression sont accompagnées d'une grande dépendance contextuelle. Une expression quelconque du langage naturel est infiniment ré-interprétable. Le sens qui lui sera attribué par le lecteur ou l'auditeur dépendra de son éducation, de sa culture, des circonstances de

lecture ou d'audition, de ce qu'il sait de l'auteur, etc. A la limite, il y a autant de lectures possibles d'une même expression que d'humains sur terre. Et, pour un humain donné, il y aura, à la limite, autant de lectures possibles d'une même expression que d'instants dans la vie. L'accord sur le sens des expressions du langage naturel, qui permet aux hommes de communiquer (ou, plus exactement, d'avoir le sentiment de communiquer) est le résultat d'une négociation de tous les instants¹. Cette négociation se produit en permanence au cours de chacune des activités pratiques. Et le sens qui est attribué aux mots, même s'il peut être grossièrement et temporairement figé dans les dictionnaires, varie en permanence avec des amplitudes considérables.

La création et l'utilisation des formalismes artificiels visent justement à éviter au maximum cette infinitude potentielle des interprétations. Les expressions dans les formalismes artificiels cherchent à transmettre aussi fidèlement et objectivement que possible au lecteur (homme ou machine) la connaissance de l'auteur. Pour parvenir à cela, les langages artificiels reposent sur un nombre de postulats restreints et connus de tous ; de plus toute assertion du langage artificiel qui n'est pas un postulat résulte d'une démonstration, d'une explication qui ne peuvent utiliser que des postulats ou des assertions préalablement démontrés. Toute l'activité du chercheur consiste donc à imaginer de nouvelles assertions et à prouver, dans le cadre du formalisme, leur cohérence.

1.- Cette façon de concevoir la notion de "sens" est caractéristique de l'Ecole interactionniste. Elle fait l'objet d'un large débat qui sortirait du sujet de cette thèse. Nous adoptons ici cette position pour faciliter l'exposé, et nous sommes bien conscient que des positions très opposées à celle-ci pourraient également être défendues.

C'est cette façon de présenter certaines activités de recherche qui a fait dire au logicien Bertrand Russell que "les mathématiques sont une science où l'on ne sait pas de quoi l'on parle ni si ce qu'on dit est vrai" ou encore à Léon Brillouin "le contenu informationnel des mathématiques est nul". D'un certain point de vue, l'activité principale en informatique consiste à faire exclusivement des manipulations symboliques. En effet, les langages de programmation sont absolument caractéristiques des langages artificiels car ils reposent sur un nombre limité de primitives : les instructions et commandes qui sont effectivement exécutables par la machine.

Bien évidemment se pose immédiatement le problème de savoir si, en utilisant un langage artificiel, il est possible de décrire l'ensemble du monde et également de décrire les langues naturelles elles-mêmes de façon suffisamment précise pour que des ordinateurs interprétant ces descriptions soient capables d'avoir des comportements linguistiques et cognitifs comparables à ceux des humains. La réponse actuelle à cette question est que les langages artificiels utilisés par l'informatique aujourd'hui ne permettent de décrire que des petits segments du monde et des petits segments des langues naturelles. La question reste ouverte.

En l'état actuel des principes de fonctionnement des ordinateurs, et sauf à abandonner le projet de recherche, on n'échappe cependant pas à l'obligation de décrire le monde en utilisant les langages artificiels actuellement disponibles. Les informaticiens vont donc s'efforcer de rédiger des descriptions d'actions (lire, écrire, parler, traduire, corriger, écouter...) et des descriptions des langues, l'évaluation ultime de la qualité de la rédaction étant la performance linguistique de l'ordinateur qui utilise ces descriptions. Bien entendu, rien n'interdit aux informaticiens d'aborder le problème de

différentes façons : par exemple, ils pourront tenter de décrire un processus d'apprentissage de la langue plutôt que (ou en même temps que) de décrire la langue elle-même.

1.3.3.- La linguistique et ses différentes branches

L'objet premier de l'informatique n'est pas d'étudier le langage naturel des humains. Ce but est poursuivi par d'autres disciplines, dont la première est la linguistique¹. Les linguistes cherchent à identifier les phénomènes caractéristiques du langage dans toute leur étendue et à en donner des descriptions plus ou moins formelles (c'est-à-dire utilisant plus ou moins des langages artificiels). Il s'agit pour eux de mettre en évidence les règles qui donnent une description de la structure ou des fonctions des phrases d'une langue. Ce sont ces règles qui correspondent à la compétence linguistique d'un sujet (ses capacités générales à propos du langage, ou, autrement dit, l'ensemble des possibilités qui lui sont données par le fait qu'il maîtrise la langue).

Ces possibilités sont d'ordres très variés : construire et reconnaître l'infinité des phrases correctes, interpréter celles qui possèdent un sens, percevoir que certaines phrases très différentes par le sens sont grammaticalement très proches, que des phrases de structures différentes sont très proches par le sens...

Dès le début des années soixante, alors que se développaient les premières recherches sur la traduction assistée par ordinateur (T.A.O.), le philosophe Joshua Bar-Hillel a montré

1.- DUCROT Oswald et TODOROV Tzvetan, *Dictionnaire encyclopédique des sciences du langage*, coll. Point, Ed. du Seuil.

que pour manipuler efficacement la langue naturelle, on ne pouvait seulement se contenter d'opérations sur les formes (qui sont décrites par la morphologie et la syntaxe) mais qu'il fallait aussi, pour pouvoir lever les ambiguïtés inhérentes à tout texte écrit en langue naturelle, comprendre effectivement ce qu'il y avait derrière les formes (ce qui est décrit par la sémantique et la pragmatique). L'un des plus célèbres exemples donné par Joshua Bar-Hillel est la phrase "*the pen is in the box*" qui peut avoir deux sens totalement différents. De nombreux exemples similaires existent dans toutes les langues : "*l'avocat est marron*".

Ainsi donc, pour faire manipuler la langue par des machines on a initialement cru qu'il suffisait de modéliser le vocabulaire et la syntaxe. On s'est rapidement aperçu qu'il fallait, de plus, modéliser de nombreuses autres connaissances. Non seulement des connaissances sur la langue elle-même (quels mots lui appartiennent, quelles règles de grammaire elle suit...), mais aussi des connaissances générales sur le monde, sur la culture de celui qui parle, sur la situation de communication, sur la pratique des relations humaines. Il est en effet important d'arriver à formaliser ces connaissances si l'on veut améliorer les performances des machines.

Mais quelques phénomènes linguistiques rendent très difficiles ces modélisations.

Les langues naturelles partagent toutes, sous une forme ou sous une autre, diverses caractéristiques essentielles qui impliquent généralement qu'il n'y a pas de correspondance exacte entre l'ensemble des mots ou des phrases et l'ensemble des "sens" (ou concepts). Ces caractéristiques ont pratiquement toutes été écartées des langages artificiels, ce

qui a rendu possible l'écriture des logiciels qui compilent et interprètent ces langages artificiels.

Ces caractéristiques communes à toutes les langues naturelles sont principalement la polysémie, la paraphrase et le rapport au contexte.

La polysémie est le fait qu'à un mot donné peuvent correspondre plusieurs sens distincts (homonymie) ou qu'à une forme de phrase donnée correspondent des interprétations diverses (homotaxie).

Le phénomène d'homonymie représente le cas général où la même graphie (ou le même son) peuvent correspondre à plusieurs emplois distincts et pose une des questions les plus importantes lorsque la machine s'efforcera d'interpréter un énoncé parlé ou écrit : celle des ambiguïtés.

Il est également difficile de tenir compte de cette caractéristique dans les machines qui engendrent de la langue naturelle, puisqu'il faudra disposer d'une représentation fine de l'interlocuteur afin de déterminer s'il pourra lever les ambiguïtés qu'on laissera subsister dans le discours.

Les difficultés posées par le phénomène symétrique des synonymes sont du même ordre : des termes différents ont des sens équivalents, et cette équivalence peut être pratiquement stricte ou bien introduire des nuances de divers ordres.

On pourra distinguer par exemple le degré d'intensité (*bassesse, abjection*), la différence d'affectation (*concierge, portier*), le niveau de langage (*personne, homme, mortel, quidam*), le degré de précision (*réparer, restaurer, raccommoder*)...

Il faudra savoir, selon les applications envisagées, si la considération de ces nuances est utile ou non et mettre en place une description du lexique et des procédures d'utilisation en conséquence.

La paraphrase correspond au phénomène symétrique qui fait qu'un même concept peut être énoncé de façons diverses : plusieurs mots possèdent le même sens (synonymie), des phrases différentes recouvrent la même idée (allotaxie) et des équivalences existent entre un mot et une phrase (définition).

Au niveau de l'analyse d'énoncés, l'ambiguïté pose le problème de reconnaître les différents sens possibles sous une forme donnée, d'avoir la possibilité de construire des représentations différentes permettant de distinguer chacun de ces sens, et de déterminer selon le contexte le sens qu'il convient d'adopter. Les formalismes de représentation utilisés doivent aussi permettre de traiter la paraphrase en donnant des représentations identiques pour des formes qui possèdent le même sens ou, tout au moins, en permettant qu'un mécanisme de raisonnement le découvre.

En ce qui concerne la génération de textes et de paroles, le problème qui se pose, par rapport à ces questions, est fondamentalement le problème du choix des mots et des structures qui vont permettre de traduire un sens (exprimé à l'intérieur de la machine par une expression dans un langage artificiel, résultant par exemple d'un calcul) sous une forme de surface.

Le rapport au contexte implique des références à l'ensemble du discours et à la situation du locuteur pour rechercher les antécédents des pronoms ou reconnaître ce que désigne telle

ou telle expression (anaphore), pour mettre en évidence ce qui est resté implicite (non-dit ou glissement de sens d'un mot par rapport à son sens propre, ce qu'on désigne du terme général de trope), ou pour s'adapter à l'interlocuteur, comprendre ses buts et situer ce qu'il dit par rapport à sa situation particulière au moment de l'énonciation (repérage).

Comme le souligne Joseph Mariani¹, le problème de la compréhension automatique du langage naturel pose fondamentalement au moins trois questions :

- que signifie "comprendre le langage" ?

La réponse n'est pas claire pour l'homme, mais elle paraît encore plus complexe si l'on se demande comment montrer qu'un automate a compris. Le sens d'un texte, en effet, n'est pas constitué uniquement de la juxtaposition des sens des mots puis des phrases, mais aussi du contexte dans lequel il est lu.

- Comment un logiciel peut-il reconnaître que quelque chose n'a pas de sens, ou tout au moins un sens plus ou moins clair ? Outre de nombreuses références implicites ou explicites à des connaissances que n'a pas le lecteur, l'utilisation des métaphores, la transformation du sens des mots en cours de texte, sont couramment exploités.

- Quelle est la fonction du langage ?

Parler ne se fait pas sans but et il faut voir le langage comme un moyen, parmi d'autres, pour résoudre des problèmes. Il est donc nécessaire de comprendre les buts de l'interlocuteur ou de lui en attribuer lorsqu'on ne les comprend pas.

1.- MARIANI Joseph, *Les techniques d'entrée/sortie vocales*, Encyclopédie CIMAB, tome 5, Paris, 1988.

Cet aspect du dialogue est extrêmement complexe à traiter, car il prend en compte, entre les humains, des phénomènes non verbaux, mimiques et postures en particulier. Il demande que soit défini un modèle informatisable du comportement des interlocuteurs.

Dans des domaines très restreints (quelques centaines de mots), on a construit dans les années 1975-80 des systèmes experts capables de déductions sur le contenu des textes, qu'il s'agisse de questions à des bases de données, de textes à résumer ou à traduire. Ces automates donnaient l'impression d'une véritable compréhension (analogue à la compréhension humaine) du texte sur lequel ils travaillaient, compréhension qu'ils n'avaient pas en réalité. Leur capacité d'extension (à des vocabulaires plus larges ou à des domaines plus larges) n'a pu, à ce jour, être démontrée. Les débats¹ qui ont résulté de ces constats ont profondément remis en cause une des hypothèses des premiers temps de l'intelligence artificielle, à savoir qu'il suffisait de doter les automates de l'apparence de la compréhension.

Mais s'attaquer à l'essence même des processus cognitifs, et non plus simuler leur apparence, est une tâche dont l'ampleur paraît être d'un autre ordre de grandeur. Après la période de grands espoirs éveillée par les premiers travaux en intelligence artificielle, les chercheurs les plus optimistes considèrent aujourd'hui que donner à une machine des compétences linguistiques et cognitives analogues à celles d'un enfant de cinq ans serait une immense avancée. Un nombre important d'entre eux pense que ce résultat ne peut être

1.- DREYFUS Hubert , *Intelligence artificielle, mythes et limites*, Paris, Editions Flammarion, 1984, 443 pages.

obtenu qu'en dotant la machine de capacités d'apprentissage autonome et de moyens de perception beaucoup plus importants que ce qui est réalisé aujourd'hui. Ces programmes de recherche constituent des axes clés dans les actuels développements autour des ordinateurs dits neuronaux.

1.4.- Connaître la langue pour la traiter automatiquement

Pour être en mesure de traiter automatiquement le langage naturel, la machine doit, comme nous venons de le voir, "connaître" (au moins partiellement) le monde, mais il faut aussi qu'elle "connaisse" la langue, tout comme un juriste, qui peut lire un texte traitant de son domaine de spécialité seulement parce qu'il possède, en plus de ses connaissances juridiques, des compétences sur la langue dans laquelle est rédigé le texte.

Les connaissances sur la langue, que l'on sait aujourd'hui formaliser pour les rendre utilisables par une machine, sont essentiellement des connaissances morphologiques et syntaxiques.

D'autres approches, ne nécessitant pas de connaissances linguistiques pour traiter la langue, ont été utilisées. Ainsi, certaines méthodes (principalement probabilistes et statistiques) ne cherchent pas à reconnaître directement les mots prononcés, mais utilisent des connaissances statistiques sur les occurrences de couples de phonèmes (bigrammes) ou même d'ensemble d'éléments plus nombreux (multigrammes). Les méthodes probabilistes employées dans ce cas relèvent de la théorie des processus. Des méthodes analogues ont été également employées pour traiter l'écrit : on ne cherche alors

plus à mettre en évidence les règles de la grammaire sous-jacente, mais à évaluer, sur la base de critères statistiques, la validité de telle ou telle séquence de mots.

Certains systèmes¹ utilisent principalement des méthodes d'analyse statistique pour indexer ou interroger des fichiers. Aucun traitement linguistique n'est effectué. Cela se traduit par une perte d'efficacité en interrogation ou en indexation. Mais cette perte est compensée par une quasi indépendance vis-à-vis du contenu des textes ou des langues utilisées ainsi que par une grande rapidité d'exécution (la phase de normalisation "linguistique" peut être fortement consommatrice de temps machine) et une extrême simplicité d'utilisation.

Divers critères de classement des connaissances sur la langue sont possibles : historiques, linguistiques, psychologiques ou simplement opératoires. Notre présentation reprend l'usage actuel en intelligence artificielle qui tente de combiner ces divers critères en cherchant à faire appel à des méthodes efficaces et qui correspondent à des niveaux de complexité différents (traitements purement informatiques, morphologiques, morpho-syntaxiques, sémantiques) tout en gardant une relative validité par rapport à l'idée que l'on se fait actuellement des processus linguistiques chez l'homme.

En traitement de la parole, les premières connaissances linguistiques nécessaires concernent le son, considéré non pas seulement en tant que signal acoustique mais également comme véhicule de la parole humaine. La difficulté majeure rencontrée à ce niveau est due à la variation phonétique : d'un locuteur à l'autre et aussi pour un même locuteur, la nature

1.- tel que le logiciel SPIRIT dans ses premières versions.

physique des sons change sans qu'il soit possible, à l'heure actuelle, de cerner de façon significative les critères distinctifs qui permettraient de réaliser les applications en reconnaissance de la parole "multi-locuteur". Phonétique et phonologie représentent probablement les parties les mieux formalisées de la linguistique et constituent deux facettes pratiquement indissociables dans le cadre d'élaboration d'applications concrètes.

Dans le traitement de l'écrit, les premières connaissances nécessaires sont les connaissances lexicales. Lexicologie, lexicographie et terminologie sont les trois disciplines qui traditionnellement structurent le domaine du lexique.

La lexicologie est l'étude des significations des éléments de base de la langue (par exemple, la lettre "s" est la marque du pluriel des noms...) ainsi que de leurs combinaisons.

La lexicographie est le recensement des mots et des expressions de la langue. Ce recensement aboutit à l'élaboration de dictionnaires, textes descriptifs portant sur le lexique d'une langue. La fonction dictionnaire, qui apporte l'intelligence lexicale nécessaire à l'interprétation des discours, occupe une place essentielle dans de nombreux domaines de l'activité humaine et représente la matière première indispensable aux traitements automatiques du langage. En effet, dans tout traitement linguistique, les premiers objets linguistiques qu'il faut décrire sont les mots.

La terminologie est l'étude des termes. L'objet de la terminologie est constitué par les vocabulaires liés à un domaine particulier organisé ou considéré comme organisable, défini thématiquement et socialement. La finalité de la terminologie est essentiellement technique, scientifique ou

professionnelle. Les vocabulaires terminologiques ont été développées dans des situations où l'exercice d'une activité scientifique, technique, administrative souffrait d'imprécision. L'activité de production terminologique est de plus en plus reconnue comme un aspect important de la diffusion de l'information scientifique et technique. En effet, l'existence, pour un domaine donné, de corpus de termes correspondant à la spécificité de ce domaine est une nécessité pour chercher, produire et former dans sa propre langue.

Cependant, l'étude du lexique a subi des mutations profondes liées à la nécessité d'accumuler un matériau descriptif utilisable par des automates. En effet, dans le traitement automatique des langues naturelles, la nécessité première est d'établir un lexique complet de la langue.

Outre les connaissances lexicales, les connaissances nécessaires à un logiciel qui analyse les énoncés écrits sont des connaissances morphologiques. En effet, dans la phrase initialement fournie au système, un même mot peut apparaître sous des formes diverses. Afin de reconnaître qu'il s'agit du même mot, une solution souvent utilisée consiste à reconstruire la forme de référence à chaque occurrence du mot. On utilise alors des connaissances au sujet de :

- la forme canonique d'un mot (infinitif pour un verbe, masculin singulier pour un adjectif, singulier pour un nom, etc.)
- ses racines et ses terminaisons possibles.
- les règles qui régissent leurs associations possibles.

Après avoir reconnu les formes canoniques des mots qui composent la phrase, il faut formaliser la façon dont ils sont agencés : quel est le verbe, où est son sujet, quels sont ses

compléments... On utilise pour ce faire des connaissances syntaxiques.

Ce sont des connaissances générales qui expriment des relations formelles entre des catégories de mots, indépendamment de leur contenu sémantique et des sujets qui les utilisent. Ces connaissances sont partiellement décrites dans ce que l'on appelle habituellement la grammaire d'une langue. Les descriptions grammaticales disponibles pour l'usage des humains ne sont pas directement utilisables par des machines pour deux raisons :

- elles ne sont jamais exprimées dans un formalisme (un langage artificiel de description) interprétable par un ordinateur ;
- elles ne sont pas complètes car elles contiennent des quantités considérables de sous-entendus, d'évidences formulées qu'il serait fastidieux de répéter à chaque fois que cela est nécessaire.

Les formalismes choisis pour décrire les phénomènes syntaxiques ont une influence importante sur la forme des représentations internes que la machine va se construire à partir des énoncés qui lui sont soumis. Ces représentations forment la base de tous les traitements ultérieurs, en particulier ceux qui porteront sur la sémantique de l'énoncé. Les grammaires traditionnelles ne sont, à l'évidence pas adéquates puisqu'elles ne sont pas rédigées avec le but de servir à la construction des représentations internes aux machines. Elles n'ont donc pas d'application réelle en informatique.

Il faut également noter l'existence de certains phénomènes linguistiques qui vont demander des connaissances relativement nouvelles du point de vue de l'étude traditionnelle des langues naturelles. Ainsi, l'homme étant capable "naturellement" de redresser les fautes de frappe en lecture ou les bégaiements en audition, peu d'études ont été faites à ce sujet. Par contre, dans une machine, il faudra concevoir un module dont le rôle est de corriger les erreurs de l'interlocuteur.

De même, l'homme, lorsqu'il ne comprend pas ce qu'on lui dit ou ce qu'il lit, exécute des stratégies très variées. Il faut également que la machine puisse faire face aux "non-attendus" (mots inconnus, structures syntaxiques non prévues, sens non répertoriés...). Elle ne le fera généralement pas de façon anthropomorphe. Actuellement, la majorité des réalisations de ce domaine traite les erreurs lexicales, mais des extensions sont en cours pour aborder les erreurs syntaxiques. Le problème principal est qu'une phrase syntaxiquement correcte n'est pas nécessairement une phrase compréhensible. D'autre part, une phrase analysée syntaxiquement peut encore contenir de nombreuses ambiguïtés.

En plus de connaissances générales sur le monde et des connaissances particulières au contexte de l'énoncé, il faudra élaborer une nouvelle classe de connaissances linguistiques pour lever toutes les ambiguïtés et comprendre réellement des textes. On considère habituellement que les connaissances sémantiques représentent le "sens" des mots, c'est-à-dire qu'elles explicitent les relations entre les mots et les objets, actions ou idées qu'ils désignent, ainsi que les conditions qui permettent d'évaluer si une phrase est vraie ou non.

Les chercheurs en syntaxe considèrent généralement aujourd'hui qu'une très faible partie de la langue est décrite de façon suffisamment fixe pour qu'un ordinateur puisse exploiter la description. Ainsi Maurice Gross considère que peut-être 5% de la langue française serait aujourd'hui décrite du point de vue syntaxique. Ce chiffre n'est évidemment qu'une approximation puisque nul ne connaît l'ampleur effective de la tâche à réaliser.

Les descriptions sémantiques, pour leur part, sont encore fondamentalement balbutiantes. Elles ne sont opératoires que dans des domaines extrêmement restreints et même dans ces cas là, elles ne prennent en compte qu'une très faible partie de la variété des sens possibles.

De plus, les critères syntaxiques et sémantiques ne suffisent pas à lever certaines ambiguïtés. Il faut le plus souvent des connaissances particulières au contexte d'énonciation pour choisir la bonne interprétation de cet énoncé. Ce type d'informations est également utile pour comprendre un texte dans son ensemble. Elles rendent compte des aspects implicites du sens et sont appelées connaissances pragmatiques.

En d'autres termes, les connaissances pragmatiques précisent une représentation du monde de référence qui constitue la culture commune nécessaire aux interlocuteurs pour se comprendre. Cette culture commune est caractérisée par le fait qu'elle est implicite, non-dite, c'est-à-dire qu'elle n'apparaît en aucune sorte dans les énoncés ou les dialogues. On appelle aussi cette culture le sens commun, l'allant-de-soi (bien que pour une large part, elle n'aille de soi que pour les seuls participants engagés pratiquement dans l'action de communication).

Les recherches des années soixante-dix ont montré que les connaissances pragmatiques sont très importantes dans la compréhension du langage par les humains. En effet, le langage naturel est très efficace pour transmettre rapidement de grandes quantités d'informations. La raison de cette efficacité est que le locuteur suppose que son auditeur (ou son lecteur) connaît un grand nombre de choses sur le sujet abordé. Il suffit alors de lui préciser ce sujet, les valeurs des variables et les anomalies éventuelles dans le déroulement normal des actions.

On perçoit les limites actuelles liées à l'obligation de décrire le monde à la machine, avant même qu'elle ne commence à analyser l'énoncé. Il est clair qu'une grande partie des informations pragmatiques ne peut être fournie à la machine (à supposer que l'on sache les formaliser), puisque le programmeur ne sait pas, à l'avance, quel sera le contexte d'énonciation des textes qui seront fournis à la machine. Ou, s'il le sait, c'est qu'il a déjà compris l'énoncé et ce n'est pas la peine de le faire traiter par un pauvre ordinateur qui n'a rien demandé à personne.

En d'autres termes, la pragmatique ne comprend pas seulement des éléments d'information connus "de tout le monde" (enfin de relativement tout le monde) avant que l'énoncé n'apparaisse, mais également des éléments par définition inconnus au moment de la programmation de la machine car directement liés au moment futur de l'occurrence de l'énoncé (comme dans la situation où une personne intervient dans une conversation déjà engagée sans elle). Un traitement complet de la pragmatique suppose donc que la machine soit dotée de capacités d'apprentissage. Dans un premier temps, ces capacités ne s'exerceront que sur les énoncés précédant celui

qui est analysé. Mais pour que la machine égale l'homme sur le plan linguistique, il lui faudra bien aussi devenir capable d'apprentissage en dehors du champ des énoncés (ce qui amène l'idée que la machine a une perception active et fine du monde et donc qu'elle a conscience d'exister, puisqu'elle doit alors distinguer son soi du reste de l'univers).

Aujourd'hui, les traitements au niveau pragmatique consistent, pour la machine, à analyser comment l'information apportée par l'énoncé s'intègre dans les connaissances que le programmeur lui avait déjà décrites à propos du monde de référence ou dans les connaissances que la machine s'est construite avant d'aborder l'énoncé. Par exemple, la machine réalisera des inférences afin de tenir compte du texte dans lequel la phrase apparaît, de mettre en évidence les liens qui existent avec le contexte et d'expliquer quels éléments, parmi les connaissances communes aux interlocuteurs, sont restés implicites.

Dans des applications très particulières, comme les automates qui assurent la vente de titre de transport, il est possible de représenter assez finement les phénomènes de pragmatique à partir d'analyses qualitatives et quantitatives des comportements des usagers. Par exemple, on peut supposer avec une bonne pertinence que le but d'un humain qui engage le dialogue avec une telle machine est d'acheter un titre de transport, on peut supposer également que l'humain mettra en œuvre toutes ses compétences pour arriver à ce but et que donc il est un "collaborateur" de la machine est non un "ennemi". Les deux considérations permettent de considérablement simplifier les logiciels de gestion des dialogues entre la machine et l'acheteur du titre de transport.

Les tentatives de formalisation des connaissances sémantiques et pragmatiques que nous venons d'évoquer vont servir de base à la fabrication de différents modules dans les machines. Des interpréteurs de récits permettront d'élaborer différents points de vue sur les textes traités : mise en évidence des liens entre désirs et actions des personnages, interprétation des contradictions, découverte des thèmes sous-jacents et relations entre les différents sous-thèmes, compréhension des buts et des plans des personnages...

Parmi tous les modules qu'il faudra concevoir, on peut citer, en particulier :

- des modules de gestion de personnages construisant l'image que se forge la machine des protagonistes intervenants dans un texte. Ainsi, dans une application faisant intervenir un dialogue homme-machine, une représentation de l'interlocuteur est créée dynamiquement. Ce processus utilise des règles correspondant à des connaissances pragmatiques et pondère la crédibilité des faits qu'il construit.
- des modules de gestion de plan inférant les intentions des différents personnages d'après leurs actions et des connaissances générales sur le monde. Dans le cadre d'un dialogue homme-machine, les intentions réelles de l'utilisateur sont ainsi mises en évidence avant de lui fournir une réponse.

Il faut rester attentif au fait que les deux modules brièvement décrits ci-dessus ne sont pas aujourd'hui opérationnels (et que l'on n'est même pas certain de savoir les réaliser un jour).

Une séparation quasi-totale a été historiquement faite entre, d'une part, la morphologie et la syntaxe et, d'autre part, la

sémantique et la pragmatique. C'est sur cette séparation que reposaient les espoirs initiaux en matière d'analyse automatique de la langue et qu'ont été faites les premières tentatives de construction de grammaires formelles de la langue naturelle. L'idée était alors qu'il était possible de construire une description formelle d'une langue naturelle et que des automates analyseurs permettraient, par exemple, de passer d'une langue à une autre uniquement en restant au niveau morpho-syntaxique. Ces tentatives ont débouché sur des échecs connus. Cependant les développements considérables réalisés ces vingt dernières années, en particulier avec l'intervention de moyens informatiques de plus en plus puissants, ont permis de montrer que l'analyse morpho-syntaxique des textes permettrait de résoudre de nombreux problèmes que l'on considérait, il y a encore une quinzaine d'années, comme relevant du champ de la sémantique.

Les connaissances linguistiques qui sont utilisées par les machines en analyse de textes et de phrases le sont également lorsqu'il s'agit d'engendrer des textes ou de la parole, même s'il est parfois utile de les organiser différemment¹. Fondamentalement, on pense aujourd'hui que le mécanisme de génération linguistique consiste à effectuer des choix, séparés usuellement en deux étapes : la génération profonde (détermination du contenu et de l'organisation du texte, réalisée par ce qu'on appelle le composant stratégique) et la génération de surface (passage d'une représentation interne du sens à une forme de surface correspondante, effectuée par le composant tactique). Les performances des systèmes de

1.- Les mécanismes de la génération automatique de textes sont décrits en particulier dans deux ouvrages :
BALPE Jean-Pierre, *Initiation à la génération de textes en langue naturelle*. Editions Eyrolles, Paris 1987.
DANLOS Laurence, *La génération automatique de textes*, Editions Masson, Paris, 1985.

génération d'énoncés ne permettent actuellement pas du tout la réalisation complète de ces deux étapes. Dans des contextes très particuliers, certaines parties de ces étapes sont cependant assez bien connues et décrites.

Lorsqu'il s'agit de générer automatiquement un texte, le premier point à traiter concerne la mise en place des buts généraux. L'intervention que l'on prévoit doit s'intégrer dans un plan général et il faut décider quand un acte de langage est nécessaire, quel type d'acte est approprié à la situation (faut-il informer, demander, convaincre...?). Dans des dialogues simples, ce point est traité directement au niveau de l'algorithmique du logiciel.

Le deuxième problème est alors la conceptualisation du message. Une fois le type d'acte de langage choisi, il faut déterminer le contenu à transmettre : de quel sujet parler, à quel niveau d'abstraction ou de détail ? Il faudra également sélectionner ce que la machine va expliciter et ce qui va rester implicite... (ce dernier point en particulier nécessite l'existence d'une représentation interne à la machine de l'interlocuteur, lui permettant d'avoir "une idée" de ses compétences et de ce qu'il sera capable d'inférer à partir de ce que la machine lui dit).

Après avoir choisi quoi dire, il faut déterminer comment le dire. La machine commencera par organiser les divers éléments sélectionnés aux étapes précédentes, c'est-à-dire planifier la structure du discours. A partir de l'ensemble des informations qu'elle a choisi d'expliciter, il lui faut construire la structure la plus adéquate en étudiant comment ces faits sont reliés entre eux, quelles sont leurs importances relatives..., ce qui débouche sur l'organisation du texte en paragraphes et en phrases.

Il faut ensuite formuler les éléments sélectionnés. Les problèmes spécifiques viennent ici du choix des mots utilisés pour représenter un sens donné, ainsi que des structures syntaxiques particulières, liées à un mot donné. Interviennent également des critères comme le choix de la voix (selon le contexte, mais il existe aussi d'autres contraintes), l'utilisation ou non de pronoms (ambiguïté éventuelle, utilisation d'un modèle de l'interlocuteur), la cohérence des différents choix (concordance des temps...), les types de liens entre les phrases (juxtaposition, relativisation, subordination, coordination).

1.5.- Les industries de la langue : une classification fonctionnelle et une présentation des produits

Les considérations précédentes ont clairement montré la vanité d'une prétention à l'universalité et à l'objectivité dans la description d'un domaine dans lequel est en train de se produire une rapide évolution technologique. La tentative faite ici n'a donc pour objet que de rendre visibles des images des industries de la langue, images personnelles ou partagées par des petits groupes de chercheurs, d'industriels, de fonctionnaires.

Le terme de classification qui est utilisé dans le titre de ce sous-chapitre implique habituellement des relations de types génériques et spécifiques, avec des représentations hiérarchiques des différents concepts. Il n'est évidemment pas possible d'utiliser ce mode de représentation pour décrire un domaine où s'entrecroisent des sciences, des technologies, des secteurs industriels aussi variés.

A titre d'exemple, si l'on fait simplement la liste des domaines scientifiques qui produisent des connaissances directement utilisables par les industries de la langue, on rencontre déjà un problème d'ordonnement. La langue dans laquelle va être donnée l'ordonnement joue un rôle de premier ordre, car les concepts couverts par les termes habituellement utilisés par les traducteurs ne se recouvrent pas d'une langue à l'autre.

De même, à l'intérieur d'une même langue, les structures des dispositifs de recherche d'Etats voisins peuvent amener à classer les sous-domaines des sciences de façons très diverses et toutes également valides.

On est ainsi confronté au problème classique en documentation : soit on reste à un niveau de généralité élevé et la hiérarchisation peut être acceptée par de nombreux lecteurs, y compris dans des langues différentes, mais cette classification n'apporte que peu d'information. Soit l'on tente une description fine et, immédiatement, la subjectivité du classificateur s'exprime de façon tellement importante que la description devient difficile à accepter par le lecteur ou bien éclate en autant de descriptions que de postulats initiaux.

Ce phénomène est clairement illustré par l'historique opposition entre Noam Chomsky et Richard Montagué. Ces deux célèbres chercheurs se sont efforcés de mettre en évidence les spécifications formelles des propriétés du langage naturel en tentant la construction de grammaires complètement explicites. A ce titre, une première approche pourrait classer leurs activités dans un même sous-domaine des sciences. Cependant, ce classement serait considéré immédiatement comme une trahison intellectuelle (et en serait certainement une) par l'une ou l'autre des deux écoles. En effet, la similarité

entre les deux lignes de pensées s'arrête à la forme de leur production. Montagué voyait la description des langues naturelles comme une branche des mathématiques. En conséquence, les langues naturelles sont supposées pouvoir être manipulées en utilisant les mêmes techniques de description que les langages artificiels (comme celui de la logique mathématique).

Dans ce paradigme, le problème principal est de mettre en place un outillage suffisamment puissant pour décrire les significations couvertes par le langage naturel. Et, plus généralement, aucune attention n'est accordée à la façon dont les humains manipulent la langue naturelle ou dont les enfants l'acquièrent. Dans le paradigme de Montagué, la seule façon raisonnable de décrire les significations associées aux phrases de la langue naturelle est de fournir une description des règles en vertu desquelles nous pouvons utiliser tel ou tel élément du langage pour décrire tel ou tel élément. Cette description constitue l'aspect sémantique de la grammaire.

En opposition avec ce paradigme, Chomsky considère que la linguistique est un sous-domaine de la psychologie et que toutes les descriptions des langues naturelles doivent amener à expliquer comment l'humain peut posséder une telle capacité. En particulier, ses travaux théoriques en linguistique sont directement orientés vers la réponse aux questions suivantes : Comment les humains peuvent-ils avoir la connaissance d'une seule langue alors même que chaque langue est un système d'une complexité potentiellement infinie ? Comment font les enfants pour acquérir une langue alors que leurs autres capacités sont simultanément relativement pauvres ? Comment se fait-il que l'on retrouve des caractéristiques communes dans des langues qui couvrent des univers culturels totalement différents ? Associé au

paradigme de Chomsky, il y a la supposition que tout humain voit et interprète le monde autour de lui pour se construire, en permanence, une représentation interne de ce monde. La signification des expressions du langage naturel n'est donc pas ici obtenue en caractérisant directement la relation entre ces expressions et les objets réels du monde. Le contenu sémantique d'une expression est obtenu en donnant la relation qui existe entre l'expression et la représentation interne du monde.

L'une des conséquences importantes de ces deux paradigmes opposés est que deux systèmes conceptuellement très différents doivent être utilisés pour décrire les activités de recherche dans le domaine. Décrire les activités de la branche "Montagué" en se plaçant dans le paradigme "Chomsky" ou vice versa n'a pas grand sens, outre le fait que cela nécessite de la part du lecteur une capacité de réinterprétation peu commune.

L'examen des diverses tentatives de typologie des industries de la langue montre que ces industries sont souvent décrites aussi bien en partant du matériau initial matériau (la langue naturelle) que des écoles scientifiques de départ, des techniques qu'elles utilisent, ou encore de découpage en marchés. Ainsi une typologie des industries de la langue ne peut être que "multidimensionnelle"¹.

Nous proposons de considérer ici que les industries de la langue ont pour caractéristique principale de reposer sur l'utilisation de compétences et de savoir-faire spécifiques issus des sciences du langage. On sait également que le concept d'industries de la langue recouvre tous les produits

1.- AUGER Pierre, *La notion d'industries de la langue, essai de typologie*, In : les industries de la langue : un domaine à la recherche de lui-même, Conseil de la langue française, Québec, 1990.

résultant du traitement automatique (c'est-à-dire qu'on obtient avec l'aide de l'ordinateur) de la langue écrite ou parlée. Cette proposition permet de cerner le champ des industries de la langue en procédant par exclusion : ne sont pas considérés comme appartenant au domaine des activités comme la traduction humaine, les méthodes d'enseignement classiques ou audiovisuelles ou la lexicographie manuelle. Ainsi, pour qu'il y ait industries de la langue, une double condition est requise : transformation du matériau primaire, la langue naturelle, et automatisation partielle ou totale du processus de traitement.

1.5.1.- Les outils de description des langues

En reprenant les intuitions les plus usuelles, on considèrera que le traitement des données linguistiques et le traitement de la langue correspondent à des niveaux de complexité croissante : au bas de l'échelle se trouve le traitement informatique le plus élémentaire qui, depuis le début des années soixante, aide le chercheur à dénombrer et classer les données linguistiques, l'écrivain à rédiger son texte, le bibliothécaire à classer ses fiches. Dans le domaine de notre étude, ces traitements informatiques de base permettent de fabriquer des objets (dictionnaires informatisés, dictionnaires électroniques et corpus de références) ; les autres niveaux de traitement, par ordre croissant de difficulté seront le traitement morphologique, le traitement morpho-syntaxique et le traitement sémantique. Bien entendu nous sommes conscient du caractère grossier de la partition que nous venons de faire. Elle n'a d'autre but que de faciliter l'exposé et ne prétend en aucune sorte couvrir la complexité des réalités informatiques et linguistiques rencontrées par les spécialistes de ces domaines.

Les traitements informatiques élémentaires de données linguistiques n'impliquent pas que soient formalisées dans l'automate des connaissances linguistiques : c'est le niveau du traitement de texte de base (sans correcteur ou vérificateur), simple prolongement des "éditeurs", au sens informatique du terme. C'est également le niveau de fonctionnement de la plupart des systèmes de gestion de base de données. On y pratique uniquement la manipulation de chaînes de caractères, sans aucune intervention morphologique, syntaxique ou sémantique. Les techniques informatiques actuelles permettent de réaliser très efficacement ces traitements, y compris pour de très gros volumes de données.

On a pu ainsi fabriquer, à l'instar des bases de données bibliographiques, des bases de données lexicales ou terminologiques, versions sur support magnétique ou optique des lexiques et dictionnaires usuels, que l'on désignera ici sous le nom de "dictionnaires informatisés".

1.5.1.1.- Les dictionnaires informatisés

Il est possible d'identifier trois générations de dictionnaires informatisés, tous destinés à être consultés par des humains et qui correspondent à trois niveaux qualitatifs :

- les dictionnaires informatisés qui se présentent comme de simples bases de données textuelles. On ne peut accéder à ces bases de données que comme dans un dictionnaire classique, à savoir par les mots ;
- les dictionnaires informatisés qui se présentent comme des bases de données structurées. On pourra accéder à ces bases de

données par les mots mais aussi par les constituants lexicographiques (définitions, exemples...) ;

- les dictionnaires informatisés qui se présentent comme des bases d'informations lexicographiques et linguistiques. Dans celles-ci, chaque constituant lexical du dictionnaire est rendu accessible et il est possible de parcourir la base (on dit souvent "naviguer dans la base") en sautant d'une définition à l'autre par sélection d'un mot dans une définition, par sélection d'une catégorie lexicographique, voire par sélection d'images associées à des mots.

Tous les dictionnaires contemporains sont aujourd'hui déjà archivés sur au moins un support magnétique. Il devient ainsi possible de stocker le contenu du dictionnaire sur n'importe quel support d'archivage électronique de grande capacité (disque magnétique, CD-ROM...). Une bande de photocomposition (ou "bande de compo") contient toute l'information *textuelle* du dictionnaire, plus tous les codes informatiques qui permettent de piloter la photocomposeuse : indications de changement de police de caractères, appels des corps et graisses, valeur des interlignages, etc.

..., et <SY0010> <CO006> <CH006> <IT> <DF23, F> pied. <FA> <GF400f> COUDRE
<GF4002> <SY0410> <CO007> <CH007> coudre <EL-2/8> <SY0010> <CO006> <CH006>
- <SYf240> <CO006> <CH006> kudr <SY0010> <CO006> <CH006> v.tr. ...

*Format d'une bande de compo (logiciel CS-SYS, IBM) ;
f in de l'article "cou-de-pied" et début de l'article "coudre",
extrait du Dictionnaire général Hachette.*

Il est relativement simple, d'un point de vue technique, de débarrasser cette bande des codes "parasites" qu'elle contient, pour n'en retenir que l'information textuelle et typographique pertinente.

... <EN> <GR> coudre <MA> [<PH> kudr <MA>] v. tr. <GR> [60] ...

*Bande de compo "débalisée" ;
début de l'article "coudre", extrait du Dictionnaire général Hachette.*

Il devient alors possible de stocker le texte du dictionnaire sur n'importe quel support d'archivage électronique de grande capacité — disque magnétique, CD-ROM, etc. Mais l'édition sur support électronique d'un dictionnaire de langue, dans sa forme textuelle brute, offre peu d'intérêt. En effet, un écran ordinaire ne permet de reconstituer que très maladroitement les fonctionnalités du papier. En particulier, le feuilletage rapide est difficile à simuler sur un écran alors qu'il constitue une méthode de recherche très efficace.

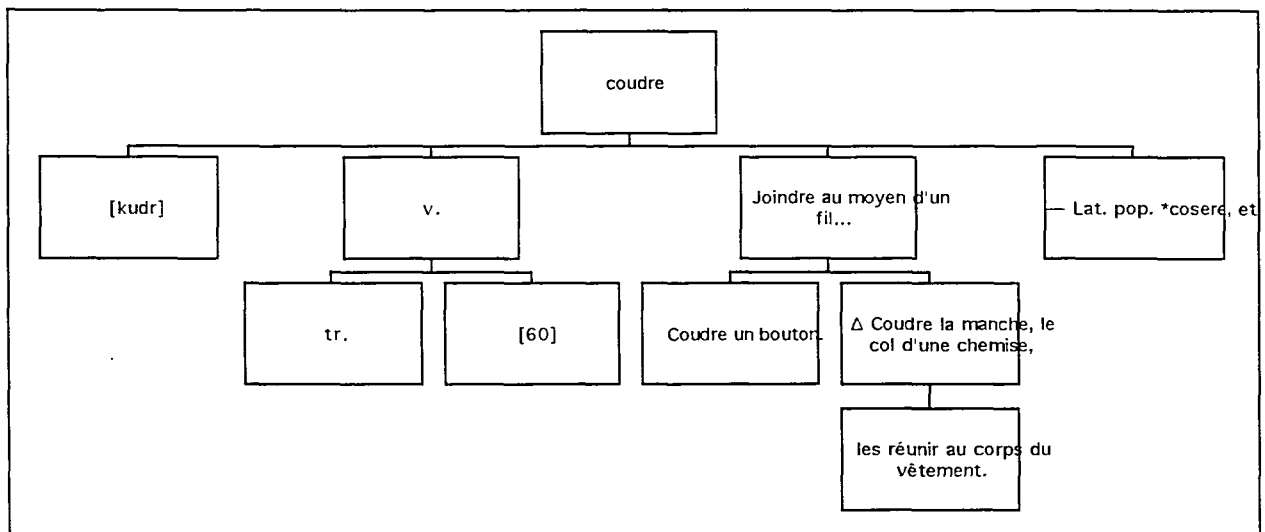
Or la réalisation d'un dictionnaire informatisé doit, pour présenter un intérêt, conduire à affranchir l'utilisateur des contraintes du papier tout en lui offrant des fonctionnalités supérieures. "Faire sortir le dictionnaire de sa *cage de papier* n'a de sens et d'intérêt que si c'est pour exploiter *l'intelligence lexicale* qu'il offre, et cela de façon telle que, par exemple, des machines puissent assumer automatiquement des fonctions de recherche lexicographique et linguistique évoluées"¹. La réalisation d'une base de données lexicale ou terminologique, à partir de dictionnaires pré-existants, n'est donc pas réduite à une simple reproduction du ou des dictionnaires sur un support différent.

Elle implique une restructuration des descriptions lexicographiques, une identification d'erreurs éventuelles et

1.- AMIEL Philippe, *Quand le dictionnaire s'automatise*, in : "Les intelligences artificielles" - revue Sciences et Avenir, n° 53 spécial, pages 94-100, 1985.

leur correction, ainsi qu'un enrichissement, tant au niveau de la macrostructure (organisation des articles) que de la microstructure initiale (structure interne des articles).

Les apports hétérogènes peuvent être "fondus", unifiés et intégrés dans une base de données qui sera contrôlée par un logiciel spécialisé dans la gestion de données linguistiques. D'une vision "gestion de texte", on évolue ainsi vers une vision "gestion de données linguistiques".

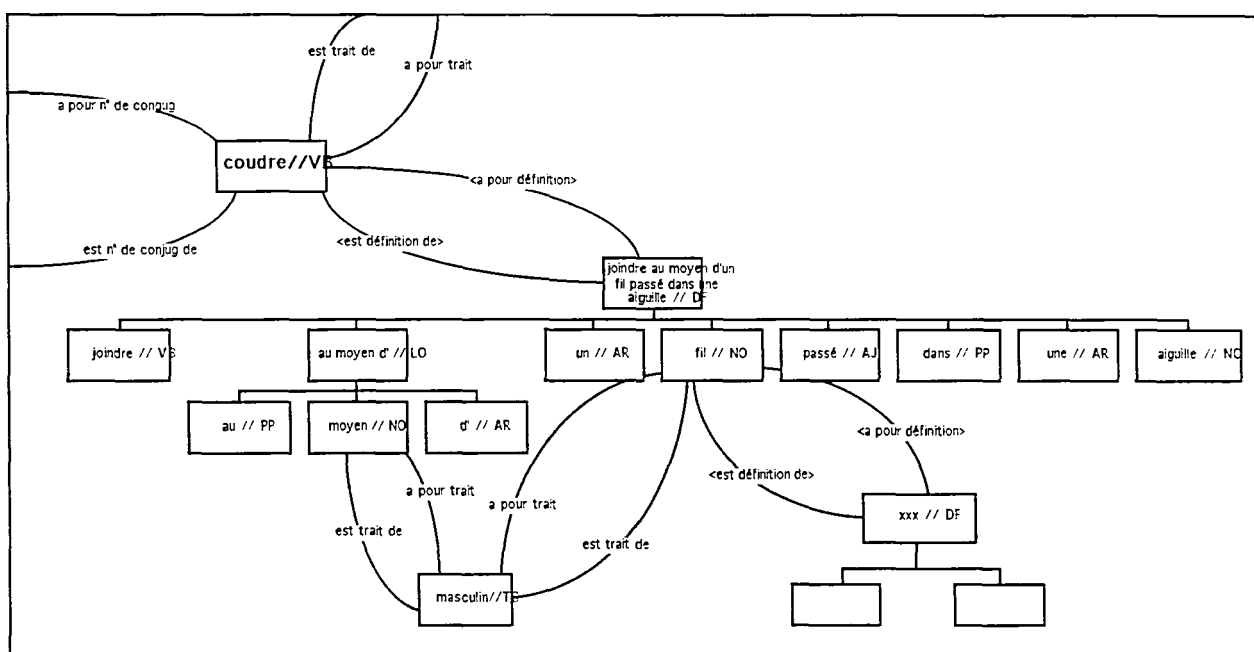


*Représentation de l'arborescence
d'un extrait de l'article "coudre" du Dictionnaire général Hachette*

La structuration des données étant relativement indépendante des exploitations particulières qui peuvent en être faites, ces bases peuvent constituer de véritables souches, permettant ainsi de développer des produits dérivés (éditions sur CD-ROM, nouvelles éditions papier, gammes de produits bureautiques...).

Bien entendu, la souche contiendra une information dont le niveau de finesse sera maximal, c'est-à-dire que les différentes dimensions linguistiques du mots seront

détaillées. Suivant le programme d'information de chacun des produits assemblés à partir de la souche, des extractions seront faites, qui ne porteront que sur une partie des entrées disponibles dans la souche ou que sur une partie des dimensions linguistiques disponibles.



*L'articulation des dimensions lexicographiques
exemple sur un extrait de l'article "coudre" du Dictionnaire général Hachette.*

Dans l'exemple ci-dessus, on peut voir comment des apports, de provenances éventuellement différentes, peuvent être rattachés les uns aux autres par un ensemble de relations qui permettront ensuite des extractions.

1.5.1.2.- Les dictionnaires électroniques

En dehors des dictionnaires à usage humain sur support informatique, il existe également une autre famille de produits destinés à décrire les langues, qui est constituée par

les dictionnaires à usage d'automates. En reprenant la proposition de Maurice Gross, nous les désignerons ici par le nom de "dictionnaires électroniques". A la différence des bases lexicales et terminologiques, consultables par des humains, les dictionnaires électroniques sont uniquement destinés à être utilisés par des automates lorsqu'ils traitent la langue, c'est à dire la résument, l'indexent, la traduisent.

Il n'existe aujourd'hui pratiquement plus d'application de traitement automatique de la langue qui ne fasse appel à l'utilisation d'un dictionnaire électronique. Ces dictionnaires prennent une importance accrue depuis quelques années. En effet, on considère généralement qu'une part importante des coûts de développement, dans de nombreuses applications de traitement des langues, sera consacrée à la fabrication des dictionnaires électroniques. Cette montée en puissance des composants linguistiques dans les systèmes de traitement automatique de la langue naturelle ne s'accompagnera pas forcément d'une disparition des techniques, essentiellement statistiques, qui étaient utilisées jusqu'à présent. On assistera cependant à un rééquilibrage qui est clairement illustré par l'évolution des vérificateurs orthographiques. Ainsi, alors ce sont des techniques statistiques qui ont d'abord été utilisées, la technique la plus utilisée aujourd'hui par les vérificateurs orthographiques est la comparaison de chacun des mots du texte à un mot d'un (ou plusieurs) dictionnaires de grandes tailles fabriqués à l'avance. Les performances de cette technique dépendent de la taille et de la qualité des dictionnaires.

Les systèmes que l'on désigne habituellement sous le nom *d'interfaces en langage naturel* utilisent également de plus en plus de données linguistiques. Ces systèmes sont intercalés entre l'utilisateur un système de gestion de base de données ou un

système expert qui, usuellement, demandent que l'utilisateur connaisse un langage de dialogue parfois peut facile d'utilisation. Les interfaces en langage naturel facilitent ainsi l'interrogation de bases de données, l'utilisation de systèmes experts ou la constitution de leur base de connaissance en permettant d'exprimer les questions sous une forme assez proche du langage humain¹. Tout comme les vérificateurs orthographiques, les interfaces en langage naturel ne peuvent fonctionner que s'ils disposent, sous une forme exploitable par la machine, de descriptions des mots que l'utilisateur va utiliser.

Une distinction essentielle existe entre les dictionnaires informatisés (destinés à l'usage par des humains) et les dictionnaires électroniques (destinés à l'usage par des machines). Les ordinateurs, contrairement aux humains, sont des machines à qui il faut "tout dire", car la machine ne dispose pas (en l'état actuel des techniques) des connaissances, des mécanismes d'apprentissage et de perceptions équivalentes ou identiques à celles de l'humain en train de lire ou de parler. Ces connaissances doivent donc être introduites préalablement, via les dictionnaires électroniques, et comme l'on ne sait pas simuler les mécanismes de l'apprentissage, il est nécessaire que ces descriptions visent à une exhaustivité que l'on ne retrouve pas dans les dictionnaires usuels. Mais, comme le souligne Maurice Gross, "tout dire suppose un préalable, qui est tout décrire et donc d'abord tout connaître".

Les dictionnaires, lexiques et grammaires usuels présentent des lacunes inévitables, parce qu'ils visent soit un public précis, soit un type particulier de description de la langue, mais aussi et surtout parce que l'humain est ainsi fait qu'il

1.- voir les exemples donnés plus loin dans cette thèse.

est capable de comprendre et manipuler la langue sans disposer de description complète. En effet, les quelques milliers de mots que contiennent les dictionnaires usuels suffisent à couvrir les besoins de communication des humains. Ainsi, une personne "normalement constituée" possède un vocabulaire de huit cents mots et un grand auteur (Proust, Balzac, Hugo...) utilise environ quatre mille mots. De même, l'ensemble du vocabulaire utilisé par une secrétaire commerciale confirmée est constitué d'environ deux mille mots¹.

Un dictionnaire classique suppose de la part de celui qui le consulte une certaine compétence. Par contre, la machine qui traite la langue n'a aucune autre connaissance a priori que celle qui est contenue dans le dictionnaire électronique.

Un dictionnaire électronique doit en conséquence être exhaustif tant du point de vue de l'ensemble des mots que de leurs variations individuelles, ceci afin de ne pas essuyer d'échec lors de la consultation par l'automate. Cette nécessité a pour conséquence d'augmenter considérablement la taille des dictionnaires et de complexifier leur structure. L'ordinateur étant une machine formelle, son fonctionnement est combinatoirement déterminé. Décrire la langue revient à en construire une description formelle. Pour désigner l'activité qui consiste à décrire formellement la plus grande part possible d'une langue naturelle, on utilise souvent l'expression "mettre à plat".

1.- Les quantités de mots qui sont données ici s'entendent pour des formes canoniques. En français, les flexions à partir des formes canoniques engendrent en moyenne huit fois plus de formes fléchies qu'il n'y avait de formes canoniques de départ.

La mise à plat du français devrait entraîner la construction d'une banque de données d'au moins cinq millions de formes. Ce nombre peut paraître très important au regard du nombre de formes que l'on peut dériver des principaux dictionnaires d'usage général (ce nombre varie de 500 000 à 700 000, suivant les dictionnaires de départ). Il faut cependant garder à l'esprit que le vocabulaire général est petit par rapport au vocabulaire technique en français, qui est de l'ordre de quatre millions de mots (presque jamais utilisés, sauf dans les domaines de spécialité).

Indépendamment de leur incomplétude, les dictionnaires usuels (sous forme papier ou informatisée) ne représentent pas une description d'un français "idéal". Ils correspondent à l'idée que le lexicographe se fait de la langue et dépendent de l'avancement des sciences et plus particulièrement des sciences du langage. L'image que les dictionnaires usuels donnent de la langue est extrêmement fluctuante. Elle est le résultat de multiples divergences culturelles, littéraires et linguistiques. Dans l'élaboration d'un dictionnaire électronique, la pression sociale ne doit plus, en théorie, s'exercer. Mais est-il "humainement" possible de dégager le lexique de tout ce qui compose notre culture et notre système cognitif dans son ensemble ? On touche ici à l'un des thèmes du débat passionnant qui oppose les linguistes syntacticiens des linguistes sémanticiens. Sans entrer en profondeur dans ce débat, remarquons quelques uns de ses points saillants :

- à la limite, chaque personne possède son propre lexique. C'est une abstraction que nous avons tous dans la tête du fait d'avoir appris une langue et de la pratiquer¹.

1.- C'est une abstraction tant que l'on en a pas trouvé la localisation et l'organisation dans le cerveau humain.

- Les dictionnaires de langue permettent aux locuteurs de se comprendre malgré les considérables différences qui semblent exister entre chaque lexique individuel ; le dictionnaire, concrétisation toujours incomplète du lexique d'une langue, comblerait ainsi les lacunes du lexique individuel.

- Le lexique d'une langue, par opposition à un ouvrage lexicographique, est un concept extrêmement abstrait et fluctuant. Le travail qu'il faudra accomplir pour construire le lexique d'une langue est essentiellement de nature anthropologique. En effet, "le mot n'est qu'une clé vers une information beaucoup plus vaste, nourrie de toute notre connaissance du monde"¹.

La nécessité de "tout dire" aux automates renouvelle la perception même des champs qui traditionnellement structurent l'étude des langues. Les frontières classiques entre lexique, grammaire et sémantique sont obligées d'éclater. L'ambiguïté "naturelle" des mots rend nécessaire l'appréhension du contexte qui, seul, permet de comprendre leur sens. Pour des chercheurs comme Maurice Gross, une manière systématique de décrire les contextes d'un mot consiste à construire sa grammaire. La grammaire devient alors un interprète des entrées lexicales plutôt qu'un recueil de faits grammaticaux très particuliers.

Ainsi, il considérera comme autant de "verbes" ce que le lexicographe considérerait comme autant de "sens" d'un même verbe.

1.- Maurice GROSS cité par Pierre CLERMONT in *le pari perdu de la traduction automatique* - Société Magazine, n°4, Juin 1990, page 27.

Par exemple, le verbe "manger", qui fait l'objet d'une entrée unique dans un dictionnaire d'usage, apparaît comme "éclaté" en plus de quatre-vingt-dix entrées. Chacune correspond à un verbe qui est considéré comme distinct des autres car tous possèdent des propriétés syntaxiques spécifiques.

Les exemples qui suivent¹ sont des phrases déclaratives simples qui comportent chacune le verbe *manger*. Les codes alphanumériques qui précèdent chaque phrase désignent, dans les dictionnaires du LADL, les classes syntaxiques auxquelles appartiennent ces verbes considérés comme distincts :

38LO	Max mange sa soupe dans un bol
35R	Max mange au restaurant
32C	La rouille mange le fer
36DT	Ce travail mange du temps à Max
32R2	La barbe de Max lui mange le visage
32L	Cet arbre mange notre vue sur la mer
32R3	Max mange son crayon
32R3	Max a mangé son héritage
32R3	Cette compagnie mange de l'argent
32R3	Ma voiture mange beaucoup d'essence
32H	Max n'a jamais mangé personne

Outre ces 10 entrées, on relève 80 formes figées (verbes et adverbes complexes), dont :

C1	Max a mangé du lion
C1	Max a mangé le morceau
PV	Max est bête à manger du foin

1.- Ces exemples proviennent du Laboratoire d'Automatique Documentaire et Linguistique (LADL) de l'Université de Paris 7.

Chaque classe syntaxique correspond à une structure syntaxique particulière. Tous les sens et nuances de sens de tous les mots doivent être systématiquement répertoriés et formalisés.

On comprend maintenant que les dictionnaires électroniques aient un nombre élevé d'entrées. Par ailleurs, l'activité d'inventaire systématique des contextes d'utilisation et des formes syntaxiques associées est extrêmement lourde. En examinant les différentes phases de fabrication des dictionnaires électroniques du LADL, on voit clairement que la fabrication des matrices qui composent ces dictionnaires électroniques demande des efforts considérables qui ne peuvent être fournis que par des humains puisque seuls ces derniers sont à même de juger de l'acceptabilité des formes rencontrées, construites ou imaginées.

Le problème des coûts de développement d'un dictionnaire électronique se pose en particulier à partir du moment où l'on envisage des applications en grandeur réelle. Le passage des maquettes de laboratoire avec utilisation des petits dictionnaires (non exhaustifs) aux applications de grande taille a permis de mesurer l'ampleur du travail exigé. Certaines de ces applications utilisent des lexiques de plus de cinquante mille mots. Le coût de rédaction du dictionnaire peut alors devenir très important par rapport au coût global du produit. On imagine donc l'intérêt qu'il y aurait à rendre les dictionnaires électroniques réutilisables d'une application à une autre.

Le problème de réutilisation des dictionnaires se pose lorsque ceux-ci ont été produits pour être utilisés dans des applications particulières. En effet, chaque type d'application (indexation automatique, traduction, génération de textes...) requiert un modèle différent et des types d'informations linguistiques différentes. Les travaux réalisés aujourd'hui résultent d'un compromis entre des besoins immédiats et la production de systèmes plus généraux. Dans plusieurs pays, des tentatives sont actuellement faites pour rationaliser la fabrication des dictionnaires électroniques en concevant des modèles normalisés de dictionnaires électroniques, qui seraient ainsi réutilisables dans des applications différentes.

En France, le projet Genelex¹ a pour objectif de construire un modèle de dictionnaires électroniques dans plusieurs langues européennes (français, italien et espagnol). Ce projet doit déboucher sur un modèle standard évolutif permettant de

1.- Projet réalisé dans le cadre d'EUREKA, qui associe des partenaires (industriels et universitaires) français, italiens et espagnols.

concevoir des dictionnaires d'application spécifiques mais compatibles entre eux.

1.5.1.3.- Les corpus de références

Le fait d'avoir pour but de "couvrir" une langue, c'est-à-dire de lui donner une large couverture lexicale et syntaxique, oblige les professionnels du domaine à accumuler de grandes quantités d'informations structurées. Celles-ci ne peuvent être classées, retrouvées, comparées et traitées que par des moyens informatiques d'une certaine ampleur.

On constitue ainsi de grands ensembles de textes écrits ou parlés, qui seront examinés en détail afin d'en extraire des dictionnaires informatisés, des dictionnaires électroniques ou des dictionnaires de sons. Nous désignerons ces ensembles par le nom de "corpus de références". L'hypothèse principale qui préside à leur réalisation est que les descriptions des langues naturelles seront plus pertinentes si elles sont réalisées à partir d'observations réelles et nombreuses de la langue plutôt qu'à partir d'une introspection des chercheurs.

Tout comme les dictionnaires, ces corpus constituent des produits nécessaires au développement du traitement automatique du langage naturel.

Construire un corpus de référence général universel pour une langue naturelle quelconque s'avère être un projet extrêmement ambitieux et aucun pays n'est encore parvenu à le réaliser. Rien ne permet d'ailleurs de dire que ce projet est réalisable puisque, la langue évoluant sans cesse, le corpus est, à tout instant, structurellement incomplet. Seuls donc des corpus limités ont été réalisés à ce jour, même si certains

d'entre eux ont pour objet une couverture large d'une langue. Leurs concepteurs s'efforcent bien entendu d'en permettre des exploitations aussi variées que possible, eu égard aux coûts importants qu'entraîne la réalisation d'un corpus.

Pour commencer, donnons quelques ordres de grandeur. Les linguistes de l'Université de Birmingham considèrent qu'un corpus permettant de réaliser un dictionnaire général de la langue anglaise doit contenir environ vingt millions de mots¹. Lorsque IBM a lancé son projet de construction d'une machine à écrire à entrée vocale, il a été décidé d'utiliser massivement des techniques de prédiction : lorsqu'un mot est reconnu, le son qui est prononcé immédiatement derrière est comparé à une table de mots phonétisés qui est ordonnancée par probabilités décroissantes d'occurrences. Si la table est bien construite, le système sera très optimisé puisque la probabilité de reconnaître le prochain mot prononcé est, en permanence, maximale dès la première recherche dans le dictionnaire des prononciations. Pour obtenir cette table, il faut examiner de nombreux textes. IBM a traité un corpus de plus de 250 millions de mots.

On peut supposer que plus grand est le corpus, plus pertinentes seront les observations faites sur le langage. Cette hypothèse doit être pondérée par le but dans lequel est constitué le corpus. Parfois, le corpus est structurellement fini. Ainsi, en est-il, par exemple, du corpus des œuvres d'un auteur décédé, ou du corpus des productions réalisées au cours d'une période passée (les poèmes épiques en français au XII^{ème} siècle). Cependant, dans le contexte qui nous intéresse, à savoir le traitement automatique et le plus généralisé possible des

¹ SINCLAIR J.M., *Lexicographic evidence*, in *Dictionaries Lexicography and Language Learning*, R. Ilson (ed.), ELT Documents, Pergamon Press, 1985.

langues naturelles modernes, on peut considérer que le corpus est potentiellement infini, puisque composé de la totalité des paroles et des écrits produits dans cette langue au cours d'une période récente. Bien entendu, dans des applications particulières, seul un segment étroit de la production totale dans une langue sera introduit dans le corpus. Ainsi, dans la machine à écrire à entrée vocale conçue par IBM, seuls des textes de lettres commerciales et techniques ont été examinés.

Si l'on cherche à réaliser un corpus très général, on remarque tout de suite qu'une partie importante du corpus d'une langue n'est tout simplement pas accessible car elle n'a pas fait l'objet d'un enregistrement au moment de sa production : il s'agit de l'immense majorité de la production orale. Les corpus de dialogues et d'échanges verbaux seront donc essentiellement constitués d'enregistrements radiophoniques ou d'enregistrements réalisés en laboratoire dans des circonstances qui ne seront plus exactement celles de la réalité (il n'y a pas de corpus de scènes de ménages enregistrées *in vivo*...).

Lors de la réalisation d'un corpus, même à visées générales, il faudra faire des choix dont on ignore le plus souvent les conséquences profondes. Ainsi, dans le corpus qui a servi à l'élaboration de la famille la plus récente des dictionnaires Collins, les choix suivants ont été faits :

- langue écrite et langue parlée
- plutôt langue générale que langue technique
- langue d'usage actuel (documents postérieurs à 1960 et, en priorité, documents très récents)
- pas de poésie ni de théâtre dramatique
- langage des adultes

- anglais standard à l'exclusion des dialectes régionaux
- principalement britannique avec quelques incursions américaines, australiennes et indiennes.

Ces critères de choix, qui relèvent apparemment du bon sens, ne sont, en réalité absolument pas évidents à sélectionner car ils peuvent fortement infléchir sans que cela soit apparent les travaux d'analyse ultérieurs. Par ailleurs, on voit que certains d'entre eux vont poser de sérieux problèmes de mise en œuvre car ils reposent sur des appréciations subjectives : comment, par exemple, déterminer qu'un texte est trop "technique" ou qu'un texte est trop "américain" pour entrer dans le corpus.

A ces critères, s'ajoutent par ailleurs des considérations qui ne relèvent pas du simple bon sens mais de choix délibérés dont les conséquences sont absolument inconnues. Ainsi, dans le corpus Collins, il a été choisi de donner priorité à des textes d'ouvrages ayant rencontré un bon succès éditorial. Cela permettait, en effet, de faire appel au réseau national des éditeurs britanniques et de faciliter la résolution des problèmes de droit d'auteur qui sont considérables à la fois par le nombre de documents manipulés, par la multiplicité des statuts rencontrés et par la variété des usages qui peuvent être faits d'un corpus.¹

Lorsque le corpus commence à être constitué, il est nécessaire d'ordonnancer les documents et donc de définir des catégories. On pourra ainsi répartir les textes écrits en

- fictions (incluant les romans historiques, les policiers, la science-fiction, etc.) ;

1.- voir, dans la seconde partie de la thèse, le chapitre qui traite de la protection juridique des corpus de références.

- réalistes (incluant les études, les guides, les thèses, etc.) ;
- lettres ;
- journaux (spécialisés et généraux) ;
- brochures commerciales et publicitaires ;

Tout comme dans des bases de données bibliographiques, les documents du corpus seront désignés par un ensemble d'informations dont certaines sont objectives (source, éditeur, année de parution...) et dont certaines seront subjectives (comme la répartition exposée ci-dessus). Les concepteurs du corpus s'efforceront de rendre les catégorisations aussi neutres que possible s'ils ne savent pas à l'avance quelles analyses seront faites. Comme la neutralité n'existe pas en la matière, ils expliciteront donc longuement le contexte de fabrication et les raisons des classifications afin de permettre aux utilisateurs futurs du corpus de contextualiser celui-ci. Par ailleurs, les bases de données seront conçues de façon à pouvoir être restructurées facilement si cela est nécessaire.

Jusqu'à aujourd'hui, ce sont les éditeurs d'ouvrages et les journaux qui disposent, sous une forme normalisée et informatisée, des ensembles de textes écrits les plus importants. Aussi, les réalisateurs de corpus ont-ils principalement fait appel à ceux-ci pour constituer les corpus élaborés dans les années 70-80. Cette situation ne va pas fondamentalement changer. Cependant, de nouvelles sources sont en train d'apparaître, grâce à la généralisation de la bureautique et de l'informatique domestique. Ainsi, il devient concevable de constituer des corpus de courriers commerciaux, de notices techniques, de textes universitaires, de notes de service dans l'administration sans avoir à ressaisir les documents originaux. Par ailleurs, les progrès considérables réalisés en matière de lecture optique

permettent de diminuer les coûts de saisie lorsqu'il n'est pas possible de disposer du document déjà entré en machine.

Les utilisations des corpus de références de textes écrits sont extrêmement variées. La plus classique est la recherche de concordances, c'est à dire l'identification de contextes d'occurrences des mots. Ces analyses permettent de fabriquer des exemples pour des dictionnaires à usage humain et de catégoriser des homographes dans des dictionnaires électroniques. Mais les corpus de références sont également très utilisés pour identifier des mots nouveaux, les équivalences entre des langues (le journal de la CEE constitue un énorme corpus multilingue du langage bureaucratique) ou les différentes graphies qui sont effectivement utilisées pour un même mot lorsque plusieurs graphies également valides sont possibles (Moyen-âge, Moyen-Age, moyen-âge, moyen âge, etc.). Enfin, les corpus de références permettent d'élaborer des tests pour tenter de mesurer les performances d'automates de manipulation du langage naturel (vérificateurs orthographiques, systèmes de TAO, etc.).

Comme nous venons de le voir, on peut trouver, dans les corpus de textes écrits, des retranscriptions de dialogues parlés. Mais il existe aussi des corpus de références qui contiennent les enregistrements proprement dits. Comme ce sont les sons qui forment l'objet des études, il n'est pas obligatoire que ces corpus soient continus (c'est à dire qu'ils contiennent des dialogues complets). On y trouvera donc des enregistrements souvent très courts (quelques secondes à quelques minutes de parole). Ces corpus de références de sons servent, d'une part, à évaluer les performances des systèmes de synthèse et reconnaissance de parole et, d'autre part, à étudier les variabilités inter et intralocuteurs en vue de modéliser la production et la perception des sons.

L'une des principales banques de sons en français est la banque du GRECO "Communication parlée" qui regroupe la majeure partie des équipes de recherche françaises sur le traitement de la parole. Cette banque de sons contient environ quarante heures d'enregistrements réalisés dans des conditions normalisées. Cette quantité peut paraître faible si on l'évalue en nombre de mots. Elle est, en réalité tout à fait considérable, car l'enregistrement est fait avec un très haut niveau de qualité, et donc consomme de grandes quantités de mémoire. Les quarante heures d'enregistrements numérisés occupent environ 4 000 méga-octets (les 20 millions de mots de la banque Collins occupent environ 800 méga-octets).

Les locuteurs enregistrés sont des locuteurs français standards : ils ont été retenus à partir de jugements sur les critères suivants : nasalité, accent, lent, clair, voix dure... Cinq juges ont retenu 12 locuteurs (six hommes et six femmes) non marqués. Par ailleurs, 10 locuteurs avec des accents régionaux ainsi que 10 locuteurs réputés difficiles à reconnaître ont été retenus. Un logiciel permet la présentation automatique au locuteur, sur écran, des éléments des corpus. Il est donc possible de travailler en permanence dans des conditions rigoureusement identiques. Un "éditeur de corpus" permet, à partir d'une simple description de l'enregistrement souhaité (conseils donnés à l'écran, phrases devant être prononcées pour enregistrement, périodes de repos...), de préparer très rapidement un programme immédiatement utilisable pour une séance d'enregistrement.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

On trouvera ci-après, à titre d'exemple, la description de deux sous-ensembles de ce corpus.¹

CORPUS "EVALUATION"

Locuteurs : 16 H et 16 F

- 1 Texte. "La bise et le soleil" -1 fois
- 2 CVCV
Avec voyelles : /a/, /i/, /u/ et consonnes : /p/, /s/, /f/, soit 54 items
- 3 Chiffres de 0 à 9. 100 chiffres.
1 liste d'apprentissage + 3 listes de reconnaissance.
4 x 100 chiffres - 1 fois
Apprentissage et liste dans le désordre.
- 4 Suites de 3 chiffres. En isolation - liste de 50 suites.
1 liste pour apprentissage (si nécessaire) + 3 listes de reconnaissance
4 x 50 suites en ordre aléatoire.
- 5 Suites de nombres de 2 chiffres (connectés) listes de 50 suites
2 fois - ordre aléatoire
- 6 Suites de nombres de 5 chiffres. Listes de 50 suites
2 fois - ordre aléatoire
- 7 Liste de 100 nombres entre 0 et 99
2 fois - ordre aléatoire
- 8 Numéros de téléphone.
100 items
- 9 Lettres de l'alphabet. Liste de 27.
(avec é et e) en isolation - dans l'ordre.
4 fois. 4 x 27 = 108.
- 10 Mots épelés en isolation. Liste de 50 noms propres.
1 fois
Le nom est donné à l'écran + présentation des lettres les unes après les autres.
- 11 Mots épelés en enchaîné. Liste de 50 noms propres.
1 fois
Horizontalement le mot enchaîné.
Verticalement ou à l'épellation - parole
a
r
o
l
e

CORPUS "ACOUSTIQUE"

1.- Cet exemple est extrait de : CARRE René, *Les banques de sons*, in Encrages Numéro 16 consacré au Colloque de Tours sur les Enjeux des industries de la langue en Europe, édité par l'Université de Paris 8, 1984.

Pour des besoins de recherche fondamentale. Les résultats peuvent être utilisés en synthèse, reconnaissance ou pour des statistiques.

- 1 Corpus CVCV.
6 H, 10 fois.
Avec les 3 voyelles cardinales /a/, /i/, /u/ et les 17 consonnes du français.
- 2 Corpus groupes consonantiques.
2 H + 2 F - 3 fois
Avec les principaux groupes consonantiques du français.
- 3 Tests de rimes (mots isolés).
Pour les consonnes par paire, par triplets ;
Pour les voyelles par paires. Total 1330 items.
- 4 Phrases phonétiquement équilibrées.
50 phrases
- 5 Phrases pour étude des nasales.
44 phrases.
- 6 192 phrases contenant des mots réels avec toutes les consonnes et toutes les voyelles du français

Après avoir considéré les outils de description des langues, dictionnaires informatisés, dictionnaires électroniques et corpus de références, nous allons maintenant essayer d'établir une typologie des produits issus des industries de la langue en nous basant sur l'étude des fonctions linguistiques qu'ils peuvent accomplir.

1.5.2.- Le traitement de la langue : une proposition de hiérarchisation des fonctions

Peut-on se contenter de la distinction classique fondée, comme nous l'avons vu, sur le niveau de difficulté du traitement linguistique ?

Dans le domaine des industries de la langue se pose un problème lié au fait que les frontières entre les sous-domaines (dans le cadre d'un même paradigme) sont extrêmement floues. L'un des exemples possibles est celui du

morceau d'arbre qui décrit la morphologie et la syntaxe comme deux sous-domaines de la linguistique, telle que la conçoit Chomsky. Tant que l'on reste dans un arbre à deux feuilles et une racine, l'accord est, à peu près, unanime. Dès que l'on commence à détailler ce qui fait partie de la morphologie et ce qui fait partie de la syntaxe, le brouillard s'installe. C'est à cause de ce flou que, de plus en plus souvent, les chercheurs limitent la distinction entre leurs activités en recherches sur la syntaxe et leurs activités en recherches sur la morphologie.

Ainsi, les dictionnaires électroniques destinés à l'usage par des automates sont souvent appelés lexiques-grammaires électroniques, exprimant non seulement le fait qu'ils contiennent à la fois des traits syntaxiques et des traits morphologiques pour décrire les entrées, mais aussi qu'il n'est pas possible de distinguer réellement les uns des autres. Lorsqu'on examine ces lexiques-grammaires il apparaît d'ailleurs qu'ils contiennent également des traits que l'on pourrait considérer comme relevant directement de la sémantique. Ainsi en est-il par exemple du trait *animé-inanimé*. Ce trait est fort utile pour distinguer les structures métaphoriques (*tuer le temps/tuer la mouche*) qui interdisent certaines transformations syntaxiques (*le temps est tué/la mouche est tuée*).

On pourrait penser que cette unification conceptuelle, impossible lorsqu'il s'agit de décrire des activités et des projets de recherche, devient possible lorsqu'on se projette plus en aval, dans l'activité de développement industriel. Les objets, ici du moins, semblent concrets et donc plus descriptibles. Il n'en est rien. Prenons l'exemple des automates de traduction qui existent vraiment (et non pas les projections pseudo-concrètes, car dans l'avenir et non dans la réalité immédiate, de théories actuellement en cours d'élaboration).

On peut tenter de faire une distinction, d'ordre très général, entre :

- d'une part les automates qui contiennent un module destiné à réaliser, sans intervention humaine, un transfert d'une langue à une autre d'un texte source vers un texte cible. Ces automates fabriquent donc un texte brut susceptible ensuite d'être révisé par l'humain (c'est, pour reprendre une dénomination un peu ancienne, la TAAH, traduction automatique assistée par l'homme) ;

- d'autre part les automates qui ne contiennent pas un tel module (c'est, pour reprendre la dénomination ancienne, la THAO, traduction humaine assistée par ordinateur). Ces automates contiennent au moins un module de gestion de dictionnaire de transfert bilingue auquel l'homme a accès pendant qu'il traduit (sinon ce sont de simples machines de traitement de texte monolingues). Et ce module contient, le plus souvent, un sous-module, dit de préparation lexicographique, qui parcourt le texte source à la recherche des mots non contenus dans le dictionnaire.

Quelle différence, du point de vue de cette distinction, peut-on faire entre SYSTRAN¹ et ce petit sous-module qui parcourt le texte et, pour chaque mot du texte source, fournit un équivalent dans la langue cible ou, *a contrario*, prévient qu'il faut faire une recherche lexicographique. Fondamentalement aucune car la seule différence est une profondeur d'analyse plus grande dans SYSTRAN, profondeur dont l'évaluation est éminemment subjective. On voit, sur ce même exemple, qu'une

1.- SYSTRAN est le système de traduction automatique actuellement le plus connu. Initialement développé pour l'US Air Force, il a été racheté par une entreprise française et il est accessible, entre autres, en mode vidéotex (3615 MITRAD).

autre distinction traditionnelle montre ses limites : celle qui était sensée exister entre les automates qui contiennent des dictionnaires utilisés exclusivement par la machine et les automates qui contiennent des dictionnaires utilisés exclusivement par l'humain. Dans la pratique, si des différences de codages peuvent exister entre les dictionnaires de telle ou telle machine, à un moment ou à un autre, l'humain comme la machine utilisent ces dictionnaires pour tenter de réaliser une forme de transfert. Ainsi donc, la séparation entre systèmes de THAO et systèmes de TAAH, qui était traditionnellement utilisée dans la littérature commerciale et industrielle, tient difficilement dès que l'on entre un peu dans le détail.

Prenons un autre exemple. On distingue habituellement les vérificateurs orthographiques (*spell checkers*) des correcteurs orthographiques (*spell correctors*). La distinction entre les deux réside dans le fait que les premiers détectent les erreurs et les signalent alors que les seconds, de plus, les corrigent. Dans la pratique, il n'y a aucune distinction possible dès que l'on regarde les fonctions réelles de ces logiciels : un vérificateur orthographique (comme celui de Word, par exemple) contient, en plus de son dictionnaire "général", un dictionnaire dit "personnel". Lorsque le vérificateur vous signale une erreur sur un mot non contenu dans le dictionnaire général et que vous la corrigez, la forme correcte est introduite dans le dictionnaire "personnel". A la prochaine occurrence de cette erreur, le vérificateur vous donne automatiquement la forme correcte.

Par ailleurs, comment qualifier les nombreux vérificateurs qui contiennent une fonction de *Recherche et Remplace* automatique ? Réciproquement, les correcteurs orthographiques actuellement disponibles ne font jamais leur

travail de remplacement d'une façon totalement automatique. En réalité, tous, lorsqu'une erreur possible est détectée, fabriquent une liste de corrections possibles triées par probabilités décroissantes et la proposent à l'utilisateur. Ainsi donc, si l'on applique à la lettre le critère de distinction, tous les produits actuels appartiennent à la même catégorie, le pur correcteur orthographique n'existant pas sur le marché.

On pourrait être tenté de distinguer les produits par leurs fonctions. Il y aurait alors d'une part les automates de traitement de texte, d'autre part, les automates de correction/vérification orthographique, par ailleurs les automates de TAO/THAO, enfin les automates d'aide à l'indexation/interrogation de bases de données, etc. Mais ici, encore, dès que l'on quitte la catégorisation générale pour examiner les fonctions réelles de ces automates, on constate que tous contiennent un module qui est un analyseur syntaxique/morphologique, que tous contiennent un dictionnaire, que tous prennent un texte en entrée pour fournir un texte en sortie, etc.

On en est donc réduit, à nouveau, à rester dans des catégorisations très générales et non disjointes : tel système sera plutôt un vérificateur, tel autre plutôt un automate d'indexation de texte, les différences entre ces systèmes ne résistant que très mal à une analyse poussée. Cette démarche de hiérarchisation nous semble cependant la seule possible dans la mesure où elle peut reposer, comme nous le verrons plus loin, sur des critères objectifs de classement.

Quant à donner les relations existantes entre des champs de la recherche et les produits réellement existants ou à venir, l'analyse des divers exercices auxquels les prospectivistes se livrent en fin de prestation dans les multiples colloques en

informatique-linguistique montre que tous les champs de la recherche peuvent intervenir dans tous les produits.

Ici encore, on peut seulement dire qu'aujourd'hui c'est plutôt le traitement du signal et la reconnaissance des formes qui permettent de fabriquer des chips de reconnaissance de la parole discontinue et plutôt les grammaires chomskiennes qui permettent de fabriquer des systèmes de TAO. Rien ne permet d'affirmer que cela continuera dans l'avenir : les ingénieurs qui travaillent sur la prochaine génération de systèmes de synthèse de la parole utilisent autant des grands lexiques-grammaires avec marqueurs prosodiques que de la synthèse de Fourier.

Comment donc appréhender le domaine des industries de la langue, puisqu'il apparaît clairement qu'il n'y a pas de concept unificateur, que ce soit sur le plan des recherches de base, de la recherche appliquée ou des produits ?

Il nous semble qu'il faut ici, modestement, utiliser les habituelles propriétés du langage naturel (ou ce que nous en connaissons). Et se contenter de multiples visions partielles, chacune d'entre elles étant contextualisée dans le temps, dans l'espace et dans un système de paradigmes aussi clairement explicité à l'avance que possible. Et donc donner au lecteur un ensemble de visions locales en attendant de lui qu'il se construise, par accumulations successives, sa propre vision du monde des industries de la langue. La liste de ces visions locales est énorme. Aussi, on n'en donnera ici que quelques-unes. Aucune ne prétendra ni à l'exhaustivité ni à l'universalité. Leur caractéristique commune est de reposer sur des faits et de s'interdire, autant que possible, toute interprétation qui soit sujette à discussion.

Il nous semble que, pour appréhender le domaine des industries de la langue, on a besoin de trois types de données de base :

- des données sur la science en train de s'élaborer (la recherche fondamentale) ;
- des données sur les produits en train d'être conçus et fabriqués (la recherche et développement) ;
- des données sur les produits et services en train de se vendre (le marché).

Les données sur la science en train de se faire sont entre autres :

- la liste des centres de recherche (dans chaque pays, dans chaque aire linguistique) qui disent faire du traitement automatique du langage naturel ;
 - la liste des universités qui disent enseigner le traitement automatique du langage naturel ;
 - la liste (ou le nombre) des chercheurs, des enseignants et des étudiants qui disent chercher, enseigner ou étudier le traitement automatique du langage naturel ;
 - la liste des projets de recherche (publics ou privés, nationaux ou internationaux) qui mentionnent explicitement le traitement automatique du langage naturel, avec, pour chacun des projets, la liste des chercheurs et des laboratoires impliqués, le montant du projet, l'origine du financement, les langues étudiées, la description du projet telle qu'elle est faite par ses participants ;
 - la liste des termes utilisés dans les publications des centres de recherche, dans les descriptions des enseignements et dans les descriptions des projets de recherche. Ces listes doivent être ordonnées par centres de recherche, ou par pays, ou par aires linguistiques (qui parle de quoi), ou par co-occurrences de termes (comment on parle) ;
-

- la liste (ou le nombre) des publications scientifiques indexées avec les termes appartenant à la liste précédente. Cette liste doit être ordonnée par pays, par aires linguistiques de localisation, par langues étudiées ;
- la liste des procédures de financement directes ou indirectes des projets de recherche, ordonnée par pays et par organismes ainsi que par langues étudiées.

Les données sur les produits en train d'être élaborés sont entre autres :

- une liste des fonctions linguistiques que peuvent réaliser des automates. Cette liste doit être hiérarchisée par niveau (subjectif) de difficulté. La liste des pages suivantes est une première approche de cette hiérarchisation. La technique permettant d'objectiver la hiérarchisation est simple :

il suffit d'ordonner les fonctions linguistiques qui correspondent à chaque niveau par le nombre de systèmes différents réellement commercialisés qui réalisent effectivement la fonction. La liste des systèmes réellement commercialisés est obtenue dans la rubrique suivante (les données sur les produits et services en train de se vendre). La liste des fonctions linguistiques est obtenue par la lecture des descriptions de ces produits. Le nombre de produits réalisant chaque fonction est obtenu soit en faisant confiance à la documentation technique du produit, soit en testant le produit, les deux nombres étant également intéressants car leur différence permet de mesurer l'attente du marché telle qu'elle est vue par les fournisseurs.

La liste ci-après n'est qu'une première ébauche. Son ordonnancement a été obtenu par une analyse de la littérature technique disponible. La difficulté est décroissante de gauche

à droite et croissante de haut en bas puis de droite à gauche. On n'a pas répété, pour limiter la longueur de la liste, tous les segments d'arbre en dessous de chaque niveau.

Un avantage de ce genre de listes est qu'elles sont relativement indépendantes de la langue dans laquelle elles sont exprimées. Cela ne veut pas dire qu'elles sont indépendantes de la langue dans laquelle sont analysées les fonctions linguistiques (d'où l'importance, comme nous le verrons plus loin, de caractériser les produits par les langues sur lesquelles ils travaillent). Ainsi, il est possible, mais cela doit être vérifié, que la fonction linguistique :

" - parler un texte écrit en codes ASCII
 -sans prosodie
 - sans liaisons entre les mots
 - avec silences réguliers entre chaque mot"

ne sera pas ordonnancée de la même façon pour le français et pour le japonais.

Liste de fonctions linguistiques que pourraient réaliser des automates

- écouter des mots prononcés par un homme
 - seulement des nombres de 0 à 9
 - prononcés, en discontinu
 - par un homme connu à l'avance
 - dans un environnement silencieux
 - les transformer en codes sténo
 - les transformer en codes ASCII
 - dans un environnement bruyant régulier (moteur)
 - dans un environnement bruyant quelconque
 - par une femme connue à l'avance
 - par un enfant connu à l'avance
 - par un seul humain connu à l'avance
 - par un seul humain inconnu à l'avance
 - par plusieurs humains inconnus à l'avance
 - prononcés en continu
 - par un homme connu
 - par une femme connue à l'avance
 - par un enfant connu à l'avance
 - par un seul humain connu à l'avance
 - par un seul humain inconnu à l'avance
 - par plusieurs humains inconnus à l'avance
 - seulement des nombres
 - seulement des nombres et des lettres
 - seulement des noms communs (ou propres)
 - n'importe quels mots d'une langue
 - n'importe quels mots de plusieurs langues
- parler un texte écrit en codes ASCII
 - sans prosodie
 - sans liaisons entre les mots
 - avec silences réguliers entre chaque mot
 - avec vocabulaire < 1000 mots
 - avec une voix d'homme
 - sans accent
 - avec accents
 - avec une voix de femme
 - avec une voix d'enfant
 - avec plusieurs voix possibles
 - avec vocabulaire complet d'une langue
 - avec silences calculés entre chaque mot
 - avec liaisons entre les mots
 - avec prosodie

**Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques**

- lire des caractères lisibles également par l'homme et les transformer en codes ASCII
 - des caractères spécialement dessinés pour la lecture optique
 - dans un seul corps
 - dans une seule grasse
 - clairement imprimés
 - mal imprimés mais chacun lisible par un humain
 - mal imprimés avec certains illisibles et reconstituables par plus de 50% des humains sachant lire
 - dans plusieurs grasses
 - dans plusieurs corps
 - plusieurs polices de caractères connues à l'avance
 - des caractères en apprenant les caractères inconnus avec l'aide d'un humain
 - des caractères en apprenant les caractères inconnus sans l'aide d'un humain

- traduire des textes écrits en codes ASCII
 - texte source dans une seule langue
 - ne contenant pas de codes typographiques
 - vocabulaire source < 2 000 formes
 - sans fautes d'orthographe
 - grammaire source < 100 règles
 - grammaire source >100 et <200 règles
 - avec fautes d'orthographe corrigeables par 50% des humains sachant lire
 - vocabulaire source > 2000 formes et <5000 formes
 - contenant des codes typographiques
 - texte source dans plusieurs langues

- corriger des textes écrits en codes ASCII
 - sans fautes de grammaire
 - texte source dans une seule langue
 - ne contenant pas de codes typographiques
 - vocabulaire source < 100 000 formes
 - corriger 10% des doubléments de lettres
 - détecter 10% des doubléments de lettres
 - détecter 50% des doubléments de lettres
 - corriger 10% de absences de lettres
 - vocabulaire source > 100 000 et < 500 000 formes
 - contenant des codes typographiques
 - avec fautes de grammaire

**Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques**

- **décrire et résumer des textes** écrits en code ASCII
 - texte source dans une seule langue
 - sans fautes d'orthographe et/ou grammaire
 - en résumant avec des mots
 - existant dans le texte source et dans un lexique pré-établi
 - lexique de mots simples
 - non hiérarchisés
 - hiérarchisés
 - lexique de mots composés
 - en construisant un lexique au fur et à mesure
 - avec tous les mots du texte source
 - avec certains mots du texte source
 - en résumant avec des segments de phrases
 - pris dans le texte source
 - non pris dans le texte source
 - avec fautes d'orthographe et/ou grammaire
 - texte source dans deux langues
- **chercher dans un ensemble de textes** ceux qui répondent à une question écrite
 - question dans un langage de requêtes
 - textes dans une seule langue
 - en prenant les textes qui contiennent les mots de la question
 - en passant à travers un langage documentaire
 - textes en plusieurs langues
 - question en langage naturel

Outre les listes de fonctions linguistiques, on s'efforcera de constituer les listes suivantes :

- la liste des produits en train de se faire, que ces produits déjà existent réellement sur le marché (un vérificateur orthographique) ou non (un téléphone traducteur). Un produit est défini comme explicitement destiné à remplir une ou plusieurs des fonctions de la liste précédente ainsi qu'à la commercialisation ;
- la liste des projets dans lequel au moins une entreprise est impliquée et qui mentionnent explicitement l'un des produits de la liste précédente. Cette liste doit mentionner, pour

chacun des projets, la liste des chercheurs et des laboratoires et entreprises impliqués, le montant et l'origine du financement, les langues étudiées, la description du projet telle qu'elle est faite par ses participants ;

- la liste des entreprises participant à l'un au moins des projets de la liste précédente. Cette liste doit mentionner, pour chaque entreprise, son activité principale, le lieu d'implantation de son siège, la liste de ses actionnaires principaux, le chiffre d'affaire dans l'activité principale et, si possible, le montant des investissements sur les projets dans le domaine des industries de la langue. Cette liste doit être ordonnée par produits recherchés, pays et aire linguistique de localisation du service de l'entreprise qui mène la recherche, pays et aire linguistique du siège social ;

- la liste des termes utilisés dans les descriptions des projets. Cette liste doit être ordonnée par entreprises et par pays, ou par aires linguistiques (qui parle de quoi), par co-occurrences de termes (comment on parle) ;

- la liste des organismes publics financeurs (et leurs budgets d'intervention), ordonnée par produits financés, par pays et aire linguistique de localisation ;

- la liste des responsables des décisions de financements publics et la liste des experts qui analysent les demandes de financement ;

- la liste des procédures de financement directes ou indirectes des projets ;

Les données sur les produits et services en train de se vendre sont entre autres :

- la liste hiérarchisée des fonctions linguistiques effectivement réalisées par les produits, telle qu'elle est définie au paragraphe précédent ;
- la liste des produits commercialisés. Cette liste doit contenir les fonctions réalisées effectivement et les fonctions annoncées, le prix du produit, le nombre d'exemplaires vendus, les langues dans lesquelles il fonctionne, le nom de l'organisme qui l'a réalisé, le nom des entreprises qui le commercialisent et, si possible, des références d'installation et des références bibliographiques ;
- la liste des entreprises qui fabriquent au moins l'un des produits de la liste précédente. Cette liste doit mentionner, pour chaque entreprise, son activité principale, le lieu d'implantation de son siège, la liste de ses actionnaires principaux, le chiffre d'affaires dans l'activité principale et, si possible, le chiffre d'affaires pour chacun des produits commercialisés. Cette liste doit être ordonnée par produits, langues traitées par chacun des produits, pays et aire linguistique du siège social ;
- la liste des services offerts dans le domaine des industries de la langue. Cette liste doit être étendue aux services susceptibles d'être automatisés. Cette liste peut contenir, pour chaque service, une évaluation du marché de ce service, par pays, par langues, par aires linguistiques ;
- la liste des entreprises qui offrent au moins l'un des services de la liste précédente. Cette liste doit mentionner, pour chaque entreprise, son activité principale, le lieu d'implantation de son siège, la liste de ses actionnaires principaux, le chiffre d'affaires dans l'activité principale et,

si possible, le chiffre d'affaires pour chacun des services commercialisés. Cette liste doit être ordonnée par services, langues traitées dans chacun des services, pays et aire linguistique du siège social ;

- la liste (ou au moins le nombre) des brevets déposés pour protéger les produits ou procédés effectivement commercialisés. Cette liste doit, si possible, contenir le nom du déposant et sa nationalité, la couverture géographique du dépôt, une description résumée du brevet, les langues auxquelles s'appliquent le produit ou le procédé breveté, la date du dépôt. Elle doit être ordonnée par pays d'origine, par déposants, par langues traitées et, surtout, par dates de dépôts ;

- la liste des termes utilisés dans les documentations commerciales. Cette liste doit être ordonnée par entreprises, ou par pays, ou par aires linguistiques (qui parle de quoi), par co-occurrences de termes (comment on parle) ;

- une liste difficile à obtenir, mais d'un intérêt considérable, est la liste pour chacun des produits, des acheteurs avec leur domaine d'activité principal, leur chiffre d'affaires et le lieu d'implantation du produit.

Bien entendu, ces multiples listes (encore incomplètes) ne peuvent toutes être fabriquées dans un délai court. On peut remarquer que, pour nombre d'entre elles, il n'est pas nécessaire de disposer d'une information exhaustive, un sondage statistique pouvant être suffisant. Cependant, il nous semble qu'à elles toutes, elles permettent de se faire une idée très précise de la structure et de l'activité du domaine des industries de la langue.

1.5.3.- Les produits issus des industries de la langue et leurs principales limites

Pour être complète, une typologie des industries de la langue doit comporter un classement de ses produits.

La littérature scientifique portant sur les industries de la langue propose plusieurs tentatives de classement selon divers types de catégorisation.

Ainsi le réseau francophone des industries de la langue propose quatre grandes catégories¹ : communication parlée ; intelligence artificielle ; traitement de l'écrit et informatique éditoriale ; analyse, classement et indexation de textes. A l'évidence, cette catégorisation n'est absolument pas pertinente car elle est incohérente. Elle ne décrit rien mais montre seulement que ses auteurs manquent tout à la fois de compétences en matière de techniques documentaires et de compétences sur le domaine qu'ils prétendent décrire.

Le comité québécois des industries de la langue² estime que trois axes sont actuellement développés : les outils de production de texte ou de discours proprement dits ; les outils d'utilisation des connaissances qui s'appliquent à des domaines extra-linguistiques ; les outils de développement de la langue, à l'usage des chercheurs et des spécialistes.

1.- Réseau francophone des industries de la langue, rapport intermédiaire, Paris, Daicadif, avril 1987.

2.- Rapport du Sous-comité des industries de la langue au Comité de coordination des contributions québécoises, Québec, Secrétariat québécois du Sommet de Québec, Ministère des relations internationales, 1987.

Comme le souligne Pierre Auger¹, professeur à l'Université de Laval (Québec), une liste des produits issus des industries de la langue est susceptible de recevoir divers classements selon les points de vue qu'on veut mettre en lumière.

Nous aborderons cette tentative de typologie en examinant les fonctions réellement exercées par les produits actuellement disponibles. Nous pensons, en effet, qu'en jetant un regard critique sur les applications qui fonctionnent effectivement, on comprend ensuite mieux les enjeux scientifiques et techniques sur lesquels repose le développement des produits.

Parce que chacun de nous pratique la langue avec une grande efficacité sans le savoir, les difficultés de son traitement automatique sont parfois mal évaluées. Cette sous-évaluation des difficultés a deux conséquences principales : on ne comprend pas pourquoi des problèmes apparemment aussi "simples" ne sont pas résolus depuis longtemps et on attribue aux machines, qui commencent à peine à balbutier, des compétences linguistiques qu'elles n'ont pas.

Les principales applications en traitement automatique des langues naturelles sont les suivantes :

- interrogation de bases de données,
- dialogue de commande (vocal ou écrit) entre des humains et des machines (du jouet à la centrale nucléaire),
- constitution et interrogation de systèmes documentaires,
- reconnaissance optique de caractères typographiés ou manuscrits,
- machines à écrire à entrée vocale,

1.- AUGER Pierre, *La notion d'industrie de la langue, essai de typologie*, In : les industries de la langue : un domaine à la recherche de lui-même, Conseil de la langue française, Québec, 1990.

- reconnaissance tactile dynamique d'écriture manuscrite et de signatures,
- vérification et correction orthographique (générale ou spécialisée),
- traduction de textes,
- contraction de textes,
- mise en forme textuelle ou discursive de données factuelles (des résultats d'analyse, des bilans d'entreprises, des horaires de train...)
- assistance à l'enseignement (des langues mais aussi d'autres matières),
- assistance aux handicapés,
- dictionnaires lexicographiques et terminologiques.

Dans certains grands domaines d'application, le traitement automatique des langues naturelles intervient parfois aujourd'hui très faiblement. Ainsi, par exemple, dans le volume total de textes tapés à la machine, la part de ceux qui sont vérifiés par des automates est très faible (moins de 1%). L'évolution du ratio volume traité par des humains/volume traité automatiquement est directement liée au progrès technique. Ainsi la généralisation des vérificateurs orthographiques sur les logiciels de traitement de texte fait lentement évoluer la situation (mais l'immense majorité des textes, dans la presse, l'édition, les bureaux, est encore corrigée manuellement).

L'imperfection du travail réalisé par les automates fait que, dans de nombreuses applications (traduction, indexation, contraction, correction de textes...), on préfère parler d'assistance au travail humain plutôt que d'automatisation complète. Et, même dans ce contexte, la pénétration des machines dans certains environnements de travail est lente pour des raisons techniques mais aussi pour des raisons

sociales. Dans d'autres applications, par contre, la part des traitements linguistiques assurés par des machines est importante et augmente rapidement. Ainsi, les grands systèmes de gestion de bases de données (SGBD) disposent de plus en plus fréquemment de modules qui s'interposent entre l'utilisateur et la machine pour aider celui-ci à formuler ses questions et pour les traduire dans le langage interne de la machine. Il en est de même pour des pans entiers du renseignement téléphonique (horloge parlante, résultats boursiers...).

Le traitement automatique des langues naturelles semble trouver aujourd'hui plus d'applications dans les activités professionnelles que dans la vie quotidienne. Cela s'explique par deux raisons principales :

- les systèmes dotés de capacités linguistiques évoluées coûtent encore trop cher, dans leur majorité, pour pouvoir être acquis par des particuliers. On note cependant que des systèmes dotés de capacités restreintes ont atteint ces dernières années les seuils de commercialisation auprès du grand public : dictionnaires électroniques multilingues de poche (abusivement appelés "traductrices de poche"), modules de synthèse de la parole dans les jouets, l'électro-ménager ou l'automobile, vérificateurs orthographiques et modules d'interrogation de petites bases de données dans des logiciels pour ordinateurs personnels...

- les systèmes de traitement de la langue naturelle imposent souvent des contraintes à l'utilisateur qui sont incompatibles avec un usage "grand public" et qui ne se justifient que par la recherche d'amélioration de la productivité en milieu professionnel. Ainsi, les systèmes de traduction assistée par ordinateur demandent actuellement à leurs utilisateurs des

compétences non négligeables (mise en forme typographique particulière des textes, rédaction contrainte, enrichissement du dictionnaire, etc.). De même les actuels systèmes de reconnaissance de la parole imposent souvent une élocution particulière, le port d'un casque/micro, une ambiance silencieuse, un apprentissage préalable de la voix du locuteur par la machine, etc.

Les descriptions qui suivent reprennent les principales fonctions linguistiques décrites plus haut : écouter, parler, lire, traduire, corriger, décrire et chercher.

1.5.3.1.- Ecouter et parler : les machines à dicter et les dialogues de commandes

La saisie de textes par clavier demande une dextérité qui est rarement rencontrée en dehors des professionnels de la frappe (secrétariat et clavistes de l'édition). Ce problème est dramatique en matière d'édition scientifique et technique à cause de la présence massive de formules et symboles qui imposent de jongler littéralement avec les claviers, les caractères complexes étant obtenus par combinaisons de deux, voire trois, touches pressées simultanément. La solution serait de disposer de machines à écrire à entrée vocale. Les recherches nombreuses qui ont été menées depuis la fin de la seconde guerre mondiale, se sont heurtées à deux difficultés principales.

La première est que la parole est un signal extrêmement variable. Deux personnes prononcent un même mot avec des différences très importantes de rythme d'élocution, d'intonation, de fréquences des sons émis (du grave masculin à l'aiguë enfantin). Les variations sont également considérables

pour un même individu en fonction de sa santé ou de son humeur (du chant hurlé, à la voix enrouée et au chuchotement...). Ainsi, les systèmes à commandes vocales qui sont installés sur les avions de chasse doivent prendre en compte les effets importants dus au stress en situation de combat (la fameuse voix "blanche", provoquée par une crispation violente des muscles du larynx), à l'utilisation de mélanges respiratoires inhabituels, aux effets physiologiques des fortes accélérations sur le conduit vocal du pilote. Il est également important de noter, au chapitre de la variabilité du signal, l'importance de la qualité de la prise de son. Le cerveau humain est capable de filtrer, avec des performances étonnantes, les paroles émises dans un environnement bruyant (de la machine-outil au crépitement de l'imprimante de bureau, en passant par les conversations simultanées ou le passage d'une voiture de pompiers).

La seconde difficulté rencontrée est que la parole est un signal quasi continu qui renferme et transmet un message discontinu. Les mots et les phrases, qui sont séparés par des espaces et des ponctuations dans leur forme écrite, sont émis sur des rythmes complètement différents et extrêmement variables lorsqu'ils sont prononcés. Les prises de respiration, les liaisons, les variations en fréquence et en amplitude de début et de fin de phrase, les accents toniques rendent l'analyse de la parole très différente, à bien des points de vue, de l'analyse des écrits.

Il n'y a pas, à l'heure actuelle, de règles précises permettant de formaliser les connaissances qui interviennent pour décoder le signal de parole (comme les grammaires de la langue écrite par exemple).

Après les premiers essais, menés dans les années cinquante, les chercheurs pensèrent qu'avec l'arrivée des ordinateurs, dans les années soixante, le problème de la reconnaissance de la parole humaine serait rapidement résolu. Les difficultés énoncées ci-dessus furent cependant rapidement mises en évidence et, en 1969, J. Pierce dans le "Journal of Acoustical Society of America" déclara avec provocation qu'il était impossible de reconnaître la parole, que ce soit mot à mot ou son à son.

Depuis, les recherches se sont orientées dans deux grandes voies. La plus ancienne peut être qualifiée de globale et acoustique, par opposition à la seconde, que l'on désignera sous le terme d'analytique et phonétique.

La démarche globale vise à reconnaître directement des mots. Le signal perçu par l'automate est comparé aux entrées d'un dictionnaire qui contient les images acoustiques des mots. Cette technique présente l'avantage de donner de bonnes performances (on peut atteindre 99 % de mots reconnus sur des vocabulaires très petits, tels que les chiffres de 0 à 9) et de n'exiger qu'un apprentissage nul ou léger (c'est-à-dire que l'utilisateur devra prononcer, à l'avance, une ou deux fois chaque mot qu'il va utiliser, afin que la machine constitue son dictionnaire d'images acoustiques). Mais elle est limitée par deux points majeurs :

- le nombre de mots qu'il est possible de reconnaître doit rester faible (de 50 à 200 mots), car, sinon, les temps de réponses croissent très rapidement ;
- dans les systèmes indépendants du locuteur, c'est-à-dire sans apprentissage préalable de la machine, il faudra que les mots soient prononcés séparément (avec des silences d'environ 200 millisecondes entre chaque mot, ce qui est peu naturel).

Les systèmes reconnaissant des vocabulaires supérieurs à la dizaine de mots seront dépendants du locuteur et nécessiteront un apprentissage préalable. L'école dite "connexionniste" a récemment redonné un souffle nouveau à l'approche globale.

Cette démarche permet de fabriquer des automates très performants dans des contextes précis (entrée vocale de numéros de téléphone).

La démarche phonétique part d'un point de vue différent. Il s'agit, dans ce cas, de reconnaître non pas directement les mots prononcés, mais les sons élémentaires (appelés phonèmes) qui composent ces mots. En français, une trentaine de phonèmes sont utilisés, mais il faut également compter avec les sons qui composent les transitions entre deux phonèmes et que l'on appelle diphones. Une fois les sons élémentaires reconnus, l'automate tente de trouver les mots, ou les segments de phrases qui peuvent correspondre, en tenant compte du fait que certains sons peuvent avoir été mal interprétés.

Actuellement aucun système opérationnel de ce type n'est utilisé ou commercialisé, contrairement aux systèmes d'approche globale.

Les prototypes de machines à dicter existent chez plusieurs constructeurs depuis une dizaine d'années, mais leurs performances sont encore insuffisantes pour une mise en marché. Il faut, en effet :

- les rendre relativement indépendants du locuteur, pour éviter que chaque nouvel usager soit contraint d'enseigner sa voix à la machine avant de travailler ;

- les rendre relativement indépendants du bruit ambiant, pour éviter à l'usager le port d'un appareillage microphonique ;
- les rendre relativement indépendants de la vitesse de dictée pour éviter à l'usager l'obligation d'énoncer chaque mot de façon discontinue ;
- les rendre capables de reconnaître plus de mots sans perdre en performances ni en souplesse (on ne dépasse pas actuellement des vocabulaires de l'ordre de la dizaine de milliers de formes, ce qui pose de gros problèmes pour les langues où les flexions sont importantes) ;
- les rendre mieux capables de distinguer ce qui est texte à saisir de ce qui est commande de mise en page (ou hésitations, discussion avec le collègue de bureau, etc.).

Les trois premiers problèmes évoqués ci-dessus sont en bonne voie de résolution et les vingt prochaines années amèneront certainement des progrès significatifs. On arrive aujourd'hui à travailler sur des vocabulaires de l'ordre de la centaine de mots en multi-locuteur et de l'ordre de 20 000 mots en mono-locuteur. Pour ce qui est des deux derniers problèmes, ils renvoient à des connaissances sur la langue écrite et sur le discours qui font aujourd'hui l'objet de nombreuses recherches mais n'ont pas encore débouché sur des résultats exploitables.

En effet, les machines transcrivent actuellement les paroles en textes sans en comprendre le sens mais en reconnaissant des formes de signaux acoustiques et en établissant des relations entre les signaux reconnus et des formes graphiques (il n'y a pas que le français où de très nombreuses lettres ne sont pas prononcées). Pour améliorer leurs performances, on est logiquement conduit à augmenter le nombre de formes sonores et graphiques connues de la machine. Mais alors, on augmente aussi fortement la probabilité d'occurrence d'ambiguïtés, au niveau acoustique comme au niveau graphique.

L'humain, semble-t-il, lève ces ambiguïtés par le fait qu'il comprend le sens du discours, ce qui est en dehors des capacités actuelles de la machine.

L'enjeu est donc soit de pousser encore plus loin la finesse des analyses de la machine sur les formes, par exemple en traitant plus précisément la "musique de la parole" qu'on appelle la prosodie, soit de doter la machine de capacités cognitives sur le monde en général et sur le sujet dont traite le texte en particulier.

Ces deux axes sont aujourd'hui en plein développement sans qu'il soit possible de donner de délai pour des applications aussi générales que la machine à écrire à entrée vocale. En effet, le postulat de l'intelligence artificielle qui dit que les processus de pensée sont, d'une part, mécanisables et, d'autre part, simulables sur un ordinateur tels que ceux dont nous disposons actuellement, n'a fait à ce jour l'objet d'aucune démonstration. Pour donner une analogie, nous sommes en face du processus cognitif humain dans une position semblable à celle qu'auraient eu les encyclopédistes du dix-huitième siècle face à un de nos ordinateurs transporté dans leur temps : ils constateraient le résultat mais n'auraient pas les connaissances de base (en physique, chimie, électronique, science des matériaux, logique) nécessaires à la compréhension. Et il n'est pas certain qu'ils mettraient moins de deux siècles, bien qu'ils aient la machine sous les yeux, à la comprendre.

On voit que, dans un premier temps, les machines à écrire à entrée vocale ne pourront pas remplacer purement et simplement des secrétaires. Elles serviront à faire une saisie "au kilomètre" nécessairement suivie d'une phase de correction et de mise en forme. Elles sont donc plutôt destinées à venir

augmenter les performances d'une secrétaire habituelle, mais également, lorsqu'elles coûteront moins cher, à aider considérablement les créateurs originaux des textes qui sont aujourd'hui de plus en plus nombreux à faire leur saisie eux-mêmes, sur des machines de traitement de texte et sans savoir vraiment taper à la machine : hier, il n'y avait que les inspecteurs de police et les journalistes qui tapaient avec deux doigts, mais désormais tout cadre, étudiant, chercheur peut être dans ce cas.

Parmi les principales annonces des dernières années, on peut citer IBM et Kurzweil. La machine développée par IBM a été présentée pour la première fois à l'automne 1984 par Fre Jenlinek du Watson Research Center. Mono-locuteur et n'acceptant qu'une parole discontinue, le prototype a démontré une capacité de reconnaissance des mots correcte à 95 % sur un vocabulaire d'environ 2 000 mots. Ce système, tout comme ses évolutions régulièrement présentées depuis, repose sur un travail préalable gigantesque, puisqu'il s'agit de l'élaboration de tables de probabilité d'occurrences de mots lorsqu'on connaît les mots précédents : à chaque instant, la machine calcule les trois mots suivants les plus probables, ce qui limite d'autant les incertitudes de l'analyseur acoustique. Les tables ont été établies par l'analyse d'énormes corpus de textes d'entreprises (notes internes, courriers commerciaux, etc.).

Cette stratégie de conception de la machine d'IBM s'appuie sur la constatation que le vocabulaire général des entreprises est étroit (2 000 à 4 000 mots) et que, dans une branche industrielle donnée, le vocabulaire spécialisé n'est pas non plus très important (au plus quelques centaines de mots pour une grande partie des textes). Bien entendu, avec les flexions, cela représente un ensemble non négligeable, mais dont les

dimensions sont petites par rapport à l'ensemble du vocabulaire d'une langue (en français, environ 80 000 mots donnant 500 000 formes dans le vocabulaire "général" et plus de quatre millions de mots dans les domaines de spécialité). Il faut cependant des machines capables de reconnaître une forme parmi 25 000 à 50 000 en temps réel avec un bon taux de succès. Les prototypes présentés se situent actuellement en dessous de ce niveau (ils travaillent sur 15 000 à 20 000 formes). Ils imposent, bien évidemment, que le locuteur sépare chaque mot par un court silence. Cette approche mélangeant du traitement du signal et des statistiques sur les occurrences de mots est également utilisée dans des développements en cours chez les fabricants d'ordinateurs européens (par exemple le projet Polyglot d'Olivetti).

Kurzweil Applied Intelligence (Waltham, Massachussetts) a commencé à commercialiser pendant l'automne 1985 le Kurzweil Voice System KVS-3 000, qui est mono-locuteur et sait reconnaître environ 1 000 mots. Le Kurzweil Voice Writer est annoncé depuis 1986 (mais toujours pas vendu en 1992). Il est beaucoup plus ambitieux : ses concepteurs le disent capable de reconnaître entre 10 000 et 15 000 mots d'un vocabulaire adaptable à l'utilisateur, avec un taux de reconnaissance correcte d'au moins 95 %. Il pourrait prendre la dictée à la cadence de 150 mots par minute, ceux-ci devant néanmoins être prononcés avec une pause entre chacun, afin de les délimiter. Cette cadence représenterait un progrès considérable, puisqu'un usager performant au clavier tape à une vitesse moyenne de 50 mots par minute, et que l'écriture à la main (lisible...) ne dépasse pas 20 mots par minute.

Le KVV serait multi-locuteur avec apprentissage en cours de saisie. Le prix de vente pourrait, dès la première année, être inférieur à 20 000 dollars (environ 110 000 francs).

Quelles sont les méthodes utilisées par Kurzweil pour en arriver aux performances annoncées ?

D'une part, des microprocesseurs spécialisés ont été construits, en particulier pour l'analyse acoustique. Ainsi, pour le filtrage frontal, le KVV utilise 25 puces KSC 2405, montées en parallèle et qui totalisent, à elles seules, une puissance de calcul pour l'application de l'algorithme de filtrage équivalente à celle d'un millier de processeurs 68000 Motorola. Parallélisme des calculs et processeurs spécialisés servent ainsi dans la plupart des étapes de la reconnaissance acoustique. L'analyse linguistique, qui complète et affine l'analyse purement acoustique, repose sur plusieurs techniques de description de la langue. Dans le KVV, ces différents "experts" sont implémentés sur des processeurs spécialisés, chacun étant affecté à un seul type d'analyse et, par là même, très optimisé du point de vue de la vitesse de traitement. Lorsque les résultats des différents modes d'analyse mènent à des contradictions, des arbitrages par vote sont faits.

Ainsi, le KVV utilise des techniques classiques d'analyse morpho-syntaxique des phrases en cours de dictée, générant des arbres de structures habituels. Cependant la gestion des arbres résultant des analyses morpho-syntaxiques a été aménagée, afin de permettre au système de faire une analyse partielle, alors que la phrase n'a pas encore fini d'être prononcée. Et le résultat de cette analyse est utilisé par un autre "expert" pour tenter de prévoir statistiquement les prochains mots qui peuvent être prononcés. Ces entrelacements de compétences permettent de jouer pleinement sur des caractéristiques de la langue. Par exemple, en anglais, les homonymes ont, le plus souvent, des caractéristiques syntaxiques très différentes, qui peuvent

être déterminées par le contexte. Les mots *two*, *to*, *too* appartiennent à des catégories grammaticales distinctes et relativement aisées à identifier par le contexte. Ainsi l'analyse syntaxique, associée à la prédiction statistique, vient au secours des processeurs de reconnaissance acoustique qui travaillent sur la base d'un fond de 30 000 formes de mots, constamment remis à jour par élimination des formes les moins utilisées.

Aucune machine n'a encore fait ses preuves en situation réelle d'utilisation. Même si les taux de reconnaissance annoncés, généralement entre 95 % et 97 % des mots dictés, peuvent sembler élevés, il faut les ramener au nombre d'erreurs par page de texte entrée. Ces taux représentent environ 15 mots faux par feuillet, ce qui est du niveau d'une mauvaise secrétaire. Les limitations sur les vocabulaires sont, par ailleurs, encore trop grandes et les machines ne fonctionnent pas dans l'environnement normal d'un bureau.

Une application intermédiaire entre la situation actuelle et les machines à écrire à entrée vocale consiste à traduire des textes sténotypés (saisis non pas manuellement mais avec des machines à sténotyper) en écriture correctement orthographiée. Le prototype le plus connu est celui développé par IBM France. La sténotypie décompose les phrases en syllabes en ne respectant pas l'orthographe (on ne prend pas en compte les homonymes comme *sot* et *seau*) et en ne plaçant pas de limites entre les mots (le mot *efficacité* en sténotypie peut correspondre à *et fis cas si thé*). De plus, l'une des 21 touches du clavier spécial utilisé par les sténotypistes peut correspondre à deux phonèmes différents. On a ainsi des ambiguïtés du type *porc* et *pôle*. A cause de ces caractéristiques, il y a en moyenne de deux à cinq formes possibles qui correspondent à une suite de caractères de

sténotypie. Le taux d'erreur des systèmes de traduction qui fonctionnent actuellement est de l'ordre de 5% et correspond généralement à des fautes d'accord. Le marché de ces systèmes est limité par l'utilisation peu répandue de la sténotypie.

Les systèmes de reconnaissance de la parole, capables de reconnaître ce qui a été dit ou de reconnaître la personne qui a parlé, peuvent être couplés aux systèmes de synthèse de la parole et permettent ainsi de réaliser des applications telles que les dialogues de commande.

Le dialogue de commande en milieu professionnel entre des humains et des machines va nécessiter les mêmes techniques que celles mises en œuvre dans les applications d'interrogation de bases de données : reconnaissance de la parole de l'opérateur, gestion de dialogues, analyses linguistiques sur la morphologie, la syntaxe et la sémantique des énoncés avec création d'une représentation interne à la machine des questions ou instructions de l'opérateur humain, traitement de ces instructions en fonction de la tâche à accomplir, génération des séquences de réponse de la machine à l'humain, synthèse de la parole à partir de texte.

Cependant, la diversité des circonstances fait que des difficultés complémentaires vont être rencontrées, par exemple le travail en milieu très bruyant, mais aussi que de nombreuses applications ne nécessiteront pas l'implémentation de la chaîne complète de dialogue. Ainsi, des cartes électroniques de transformation texte-parole permettent la synthèse de phrases quelconques (le message à synthétiser est représenté sous forme de graphèmes) et donnent une voix synthétique intelligible. Cette voix sera suffisante pour que la machine attire l'attention de

l'opérateur, lui dicte des instructions au fur et à mesure qu'il exécute la tâche. Le fait que la voix soit métallique, ou sans prosodie, n'est pas gênant car cela n'empêche en rien la transmission d'informations de la machine à l'homme. Les utilisateurs professionnels apprécient souvent les caractéristiques très stables d'une même voix de synthèse qui devient vite facile à reconnaître et à comprendre.

De même qu'en milieu professionnel, de très nombreuses applications grand public de communication entre humain et machine n'utilisent que des segments de la chaîne du dialogue. A cause des bonnes performances obtenues aujourd'hui, la synthèse de voix par la machine est le segment le plus couramment utilisé. Plus personne ne fait attention aujourd'hui aux flippers installés dans les cafés qui invitent à jouer ou commentent les parties d'une voix caverneuse. Pour rester dans le champ du jouet, l'exemple le plus célèbre d'application grand public est certainement le "Speak and Spell" de Texas Instrument. Commercialisé en France depuis 1978 sous le nom de "Dictée magique", ce petit appareil est destiné à apprendre l'orthographe de mots simples. Il guide l'enfant et donne les résultats du jeu. La synthèse de parole, dans ces jouets, est réalisée par concaténation de segments sonores préalablement numérisés. La première version du Speak and Spell vendue en France avait un effroyable accent texan car les tables de segments sonores étaient les mêmes que celles utilisées dans la version américaine. Cela s'est amélioré depuis...

Il n'y a pas que les jouets qui parlent. Au Japon, Matsushita commercialise des distributeurs de boissons qui annoncent l'état de leurs stocks (thé, chocolat sucre,...) et qui interpellent le client grâce à un détecteur de proximité. Certaines automobiles donnent au conducteur des informations

parlées (ceinture de sécurité non bouclée...), de même que quelques appareils ménagers (aspirateurs, fours électriques, machines à laver...). Dans de nombreux cas, les applications gadgets n'apportent pas un véritable progrès et ne constituent qu'un appât commercial supplémentaire. La qualité de la voix de synthèse a fait des progrès considérables ces dernières années. Il est aujourd'hui possible de générer des voix d'enfants, de commentateur de football, de monstre, d'hôtesse de l'air, de guide affable. Le caractère artificiel de la voix est cependant maintenu dans de nombreuses applications afin de ne pas induire l'humain en erreur. Il est en effet souvent nécessaire que ce dernier sache qu'il est en train d'entendre une machine et non un autre humain.

1.5.3.2.- Lire et corriger : vérification et correction orthographique ou stylistique

Dans le travail d'un rédacteur ou d'un responsable de publication, la correction orthographique et stylistique constitue une part importante du temps de travail. Ainsi, les coûts de la correction dans les métiers de la presse et de l'édition sont de l'ordre de 30 % des coûts de saisie. De même, la ré-écriture stylistique représente environ la moitié du temps de rédaction dans l'édition, le tiers dans la presse mensuelle, le cinquième dans la presse quotidienne. On comprend que l'on ait très tôt cherché à automatiser ces fonctions, d'autant que le développement rapide de la bureautique constituait dès le début des années 1980 un second marché important.

Le problème que tente de résoudre un vérificateur orthographique s'énonce simplement de la façon suivante : étant donné un texte, le parcourir et identifier les mots qui

sont incorrects. On distingue habituellement le vérificateur orthographique (en anglais, *spelling checker*) du correcteur orthographique (en anglais, *spelling corrector*) qui, de plus, s'efforce de trouver, pour chaque mot incorrect, un remplaçant correct.

Les systèmes de vérification orthographique, généralement intégrés dans des logiciels de traitement de textes, fonctionnent sur la base de deux techniques différentes. La plus ancienne a été utilisée dans TYPO, un vérificateur qui fonctionne sous UNIX. L'idée est que les séquences de deux et trois caractères réellement utilisés dans l'écriture d'une langue sont beaucoup moins nombreuses que la combinatoire sur deux et trois caractères de l'alphabet. De plus chaque doublet et triplet existant réellement peut être, par analyse d'un grand corpus, affecté d'un coefficient de probabilité d'occurrence. L'analyse du texte à corriger consiste à détecter les doubles et triplets dont l'occurrence est anormale, l'expérience indiquant que les mots contenant des fautes de frappe (inversions, erreurs, omissions et redoublements de lettres qui représentent 80 % des fautes) auront un fort indice d'anormalité. Afin d'améliorer la performance du système, TYPO contient un dictionnaire des 2 500 mots les plus courants de la langue. Comme l'expérience montre que sur un texte en anglais (non littéraire) de 10 000 mots, on ne compte, en moyenne, que 1 500 mots différents, une comparaison entre la liste des mots jugés statistiquement anormaux et le dictionnaire réduit considérablement la liste des "erreurs" proposées à l'utilisateur. Cette technique, plus mathématique que linguistique, est très caractéristique des produits qui ont été réalisés entre 1960 et 1980. Ses limites en sont évidentes même si elle donne parfois de meilleurs résultats que des techniques linguistiques et qu'elle est très facile à transférer d'une langue à une autre.

La technique la plus utilisée aujourd'hui est la comparaison de chacun des mots du texte à un mot d'un (ou plusieurs) dictionnaire de grande taille fabriqué à l'avance. Les performances de cette technique dépendent pratiquement de la taille et de la qualité du dictionnaire. Deux dictionnaires sont habituellement utilisés : un dictionnaire général fourni par le vendeur du logiciel et un dictionnaire personnel contenant le vocabulaire souvent utilisé par l'utilisateur (et créé par ce dernier). Parfois, comme dans le vérificateur orthographique de WORD, un troisième dictionnaire, également créé par l'utilisateur (et appelé "dictionnaire du document"), contient le vocabulaire exclusivement utilisé dans le texte en cours de correction (afin de ne pas surcharger les deux dictionnaires précédents et de prendre en compte des cas particuliers : fautes volontaires, noms propres, etc.). D'autres structures de dictionnaires sont également utilisées. Par exemple, un premier dictionnaire, en mémoire centrale, contient les formes (de 2 000 à 5 000) statistiquement les plus utilisées dans la langue. Un second dictionnaire, également en mémoire centrale, contient le vocabulaire souvent utilisé par l'utilisateur et un troisième dictionnaire, le plus complet possible, est placé sur disque externe.

Les dictionnaires de l'utilisateur sont de simples listes de mots. Le dictionnaire général a une forme un peu plus complexe (dans les vérificateurs les plus performants, car pour les autres, il ne s'agit généralement que d'une liste de mots...). Il contient le plus souvent de 200 000 à 250 000 formes. Nous avons vu que la langue française générale comprend plus de 500 000 formes dérivées (par conjugaisons et flexions diverses) de plus de 80 000 mots de base. La prise en compte des termes techniques aboutirait à un ensemble de plus de 4 millions de

mots. La complexité du dictionnaire général est de plusieurs ordres :

- le dictionnaire ne contient pas tous les mots entiers mais des racines et des affixes, avec des algorithmes de génération, ce qui a le double avantage de fortement réduire l'espace mémoire occupé (les mots *durée*, *durable*, *durablement* sont stockés sous une forme du type *dur-ée,-able,-ment*) et de permettre, sous certaines conditions d'accepter des néologismes locaux (pratique courante en anglais). Cette compression est limitée par les multiples exceptions qui ne sont souvent gérables que par énumération (par exemple, l'accent change de grave en aigu entre *misère* et *misérable*).

- le dictionnaire ne contient pas seulement des mots et portions de mots mais aussi quelques marqueurs morphologiques et parfois syntaxiques qui servent à activer des algorithmes liés à certaines formes (par exemple, les conjugaisons des verbes, les pluriels réguliers, etc.).

- le dictionnaire peut être réparti en plusieurs sous-ensembles ayant des structures très différentes, chacune correspondant à un type de traitement : un sous-ensemble contiendra les formes phonétisées des mots, un autre contiendra les mots retournés (première lettre=dernière, deuxième=avant-dernière, etc.), un troisième contiendra les mots avec des comptages des lettres qui les constituent. Ces différents sous-ensembles servent à des algorithmes destinés à proposer des corrections pour un mot erroné qui a été détecté.

Les vérificateurs actuellement commercialisés utilisent les dictionnaires d'une façon qui n'a probablement rien à voir avec les mécanismes de mémorisation et de contrôle

orthographique chez l'humain. En effet, ces vérificateurs parcourent le texte à la recherche des caractères considérés comme séparateurs de mots (espaces blancs, virgules, apostrophes...). Ce sont les suites de lettres comprises entre les séparateurs qui sont comparées, une à une, aux formes contenues ou dérivées dans les dictionnaires.

On voit facilement que cette méthode ne couvre que très grossièrement la diversité linguistique :

- les compétences morphologiques sont limitées aux termes simples. Ne sont pas détectées, par exemple, les erreurs sur les mots composés qui contiennent des espaces blancs, tel que "*chêne de télévision*" non plus que toutes les erreurs qui transforment un mot en un autre mot existant.
- le problème de l'identification des séparateurs n'est généralement pas bien résolu, car les multiples cas particuliers ne sont pas correctement énumérés (par exemple, l'apostrophe n'est pas un séparateur dans "aujourd'hui", il y a aussi des points dans les sigles)
- le problème des noms propres amène parfois certains vérificateurs à purement ignorer tous les mots contenant une majuscule, y compris les premiers mots de chaque phrase, etc.

Les compétences syntaxiques et, a fortiori, sémantiques sont quasiment nulles, bien que des progrès aient été faits avec la génération de vérificateurs apparue au début de 1990. Ces derniers sont capables de corriger quelques accords de participes. Mais ils ne savent pas analyser l'immense majorité des fautes d'accord, de temps, non plus que toutes erreurs faisant un tant soit peu intervenir des connaissances générales sur le monde ou sur le sujet dont parle le texte.

On voit ici poindre le même problème que dans la machine à écrire à entrée vocale. Et tant que ce problème n'est pas résolu on n'assistera qu'à des améliorations de détail (mais néanmoins agréables à l'utilisateur) : accroissement de la vitesse de vérification, prise en compte de certaines expressions composées, élargissement des dictionnaires, etc.

On peut se demander à quoi peuvent donc servir des vérificateurs aussi limités ?

Ils détectent principalement les fautes de frappe : inversions, omissions, erreur et redoublement de lettres. Ces fautes sont les plus courantes chez les usagers non professionnels de claviers. Et la quasi inefficacité des actuels vérificateurs sur les "vraies" fautes d'orthographe explique qu'ils soient peu utilisés par les professionnels de la correction : les correcteurs humains professionnels considèrent que les fautes de frappe "sautent aux yeux" et qu'ils n'ont pas besoin d'un système qui attire leur attention dessus, car ce mode de travail conduit l'humain à ne plus faire attention aux autres fautes (à cause de la segmentation de la lecture-compréhension induite par les interruptions de la machine).

Evidemment, les fabricants de vérificateurs orthographiques se sont efforcés d'améliorer les performances de leurs systèmes. Mais dès que la simplicité des vérificateurs de base est abandonnée, de grandes difficultés sont rencontrées. Un exemple classique de ces difficultés est l'introduction dans l'automate d'une nouvelle règle orthographique qui augmente les fausses erreurs. Mais, dès que l'on tente d'affiner la règle, on s'aperçoit que personne ne la connaît vraiment, qu'elle n'est pas suffisamment formalisée pour une implantation sur une machine et que le concepteur du vérificateur doit faire lui-

même le travail considérable de recherche des multiples situations linguistiques possibles. S'il fait ce travail, il s'aperçoit rapidement que le nombre de règles nouvelles à introduire croît de façon considérable, augmentant les différents paramètres du temps de traitement. L'un des aspects principaux de cette spirale catastrophique est la croissance du dictionnaire qui atteint vite des volumes incompatibles avec :

- les algorithmes d'accès aux données ;
- les espaces disponibles en mémoire centrale : les importants facteurs de compression obtenus pour les dictionnaires dans les vérificateurs simples se réduisent rapidement devant l'avalanche des cas particuliers. Les dictionnaires ne peuvent plus résider en mémoire centrale, ce qui est encore possible dans les vérificateurs actuels (dont les dictionnaires compressés occupent entre 250 et 800 Ko) ;
- la technologie des mémoires de masse (les disques les plus courants sur micro-ordinateurs demandent environ 13 millisecondes par accès et il faut plusieurs accès pour un seul mot...).

Outre les effets énumérés ci-dessus, l'accroissement des dictionnaires a d'autres effets négatifs. Par exemple, la fonction qui propose des mots possibles en remplacement d'un mot identifié comme éventuellement erroné perd rapidement toute efficacité car, le dictionnaire croissant, le temps de recherche des mots-candidats devient extrêmement long. Cette fonction devient donc optionnelle dans le système et n'est, en pratique, pas utilisée. Une variante consiste à considérer arbitrairement que les deux premières lettres du mot identifié comme faux sont correctes, de façon à limiter la

recherche de candidats-mots remplaçants, mais on voit bien qu'en s'engageant sur cette voie, le fabricant du vérificateur (par exemple, Spell sur les systèmes DEC) a été contraint à un "bricolage" de mauvais aloi.

On pourrait penser que l'accroissement des tailles mémoire et des vitesses de calcul dû au progrès est suffisant pour maintenir des performances constantes lorsqu'on augmente la taille des dictionnaires. Il n'en est rien car l'explosion combinatoire linguistique est plus rapide que cette amélioration des performances techniques. Les concepteurs de vérificateurs orthographiques sont donc obligés d'améliorer sans cesse leurs algorithmes de traitement. Ces améliorations visent en particulier à limiter au maximum les listes énumératives au profit d'automates finis. Il est également fait appel à des techniques de programmation issues de l'intelligence artificielle. Quelques vérificateurs commercialisés ces deux dernières années contiennent des analyseurs syntaxiques restreints, capables par exemple de traiter certains accords de participes. On considère que ces vérificateurs détectent actuellement environ la moitié des fautes dans un texte (le nombre de fautes étant, bien évidemment, très dépendant du rédacteur et du claviste).

Si les vérificateurs orthographiques à vocation générale sont les plus connus, il existe aussi de nombreuses applications faisant appel à des vérificateurs dans des contextes linguistiques plus étroits. Tous les systèmes d'interrogation de bases de données en langage naturel évoqués dans les paragraphes précédents contiennent un vérificateur placé en amont des analyseurs et destiné à nettoyer les textes entrés par les humains avant traitement. On utilise aussi des vérificateurs limités pour contrôler les fichiers d'adresses postales avant l'envoi de publi-postages ou pour contrôler les

noms des rues et des localités dans les systèmes d'informations géographiques. Dans de nombreuses applications, il est non seulement possible de faire une vérification avec d'excellents taux de détection, mais également de corriger automatiquement une grande partie des erreurs.

La correction automatique du style d'un texte est aujourd'hui encore très peu développée. Elle s'appuie le plus souvent sur le calcul du nombre de répétitions d'un mot dans une phrase, sur la longueur d'une phrase ou d'un paragraphe. Ces petits modules sont intégrés dans certains vérificateurs orthographiques. Ils sont bien loin de savoir identifier les habituelles formes stylistiques et rhétoriques qui nécessitent une compréhension effective de ce qui est écrit...

Les linguistes ne savent pas bien définir ce qui fait qu'un texte peut être considéré comme bien écrit. Ceci d'autant que les critères de décision varient fortement d'une langue à l'autre. Ainsi, la répétition d'un même terme plusieurs fois dans une phrase est considérée en français comme une faute (sauf cas rares) et elle est parfaitement acceptable en anglais. Ces critères varient également fortement suivant le genre de texte dans une même langue. Sans même parler de littérature, on n'écrit pas une bonne lettre commerciale de la même façon qu'une bonne démonstration en mathématiques. Des critères comme la longueur des phrases, le nombre de répétitions d'un même terme, la longueur des paragraphes sont utilisés par les quelques vérificateurs de style qui fonctionnent en anglo-américain.

Des progrès considérables peuvent être attendus à échéance d'une vingtaine d'années en matière de vérification stylistique de base par transfert aux logiciels de traitement de textes des

fonctionnalités qui commencent à apparaître dans les logiciels d'enseignement des langues. Par contre, les linguistes n'ont pas encore fourni les éléments qui permettraient de construire des vérificateurs stylistiques travaillant sur l'ensemble de la rhétorique des textes. Métaphore, tautologie, métonymie, image sont encore inaccessibles à la formalisation informatisée car leur reconnaissance suppose une véritable compréhension des textes.

1.5.3.3.- Lire, décrire et chercher : la constitution et l'interrogation de systèmes documentaires

Par système documentaire, on entend ici des bases de données qui contiennent des fiches bibliographiques ou des bases de données qui contiennent des textes. Rappelons qu'une fiche bibliographique est constituée de deux parties : une partie "objective" qui donne les coordonnées de l'ouvrage ou de l'article (nom de l'auteur, titre, éditeur, année de parution,...) et une partie "subjective" (résumé, mots-clés...). Cette seconde partie est dite subjective car elle est dépendante du rédacteur de la fiche bibliographique et du but informationnel poursuivi dans la base de données : un ouvrage sur la vie des fourmis ne sera pas du tout résumé de la même façon dans une base de données en sociologie humaine et dans une base de données en biologie animale, bien qu'il puisse être pertinent dans chacun des deux domaines.

Lors de la constitution des bases de données bibliographiques, le traitement de la partie objective ne pose pas de problème particulier du point de vue des industries de la langue. Les progrès attendus portent sur la diminution des temps de saisie (par exemple avec la systématisation des codes à barres) et des aides limitées (contrôle orthographique sur les noms

d'éditeurs, vérification du respect des normes catalographiques) faisant peu appel aux traitements linguistiques. En revanche, la constitution de la partie subjective est, depuis longtemps, l'objet de recherches actives. S'il n'existe actuellement aucun système capable de constituer automatiquement des résumés analogues à ceux que produirait un rédacteur humain, plusieurs techniques ont été utilisées pour tenter de diminuer le coût considérable de cette partie du travail.

En bas de l'échelle de la complexité, on trouve les techniques uniquement morphologiques, qui reposent sur le principe de la reconnaissance de formes graphiques. Lorsqu'un titre ou un résumé (rédigé par un humain) est entré dans le système, un filtre est appliqué qui va, plus ou moins automatiquement, éliminer les mots supposés vides de sens (articles, prépositions...) et constituer les listes inverses (à chaque mot non vide, appelé mot-clé, est associée la liste des ouvrages où ce mot apparaît dans le titre ou dans le résumé). La requête de l'utilisateur doit prendre la forme d'une équation booléenne utilisant les mots-clés. Cette équation constitue un filtre sur les listes inverses. Par exemple, si l'on veut les références des ouvrages parus aux Etats-Unis ou en France depuis 1980 et traitant des protéines du sang de poulet, la requête ressemblera à ceci :

PROTEINE? ET POULET? ET SANG ET (AN>=1980) ET (ORI=USA
OU ORI=FRANCE)

Les mots PROTEINE, POULET et SANG seront cherchés dans les titres et les résumés des ouvrages. Les points d'interrogation permettent de gérer grossièrement les éventuels pluriels. AN et ORI désignent les champs "date de parution" et "pays de parution" des fiches bibliographiques.

L'immense majorité des systèmes documentaires actuellement opérationnels ne fait rien de plus qu'un traitement que l'on hésite souvent à qualifier de linguistique : l'élimination des mots vides se fait par consultation d'une liste fermée, il n'y a pas de normalisation des flexions (POULET=POULETS), c'est l'utilisateur qui gère les mots composés, etc. Les améliorations apportées depuis le début des années 1980 portent sur l'ergonomie de l'interrogation : masques d'interrogation, simplification et assouplissement de la syntaxe dans les équations de recherche.

Dans quelques rares systèmes documentaires, les mots-clés choisis par le documentaliste ne sont pas choisis de façon intuitive, mais sont extraits d'un lexique ou d'un thésaurus. Dans ce cas, lors de l'interrogation, des fonctions d'extension ou de restriction peuvent être utilisées. Ainsi, par exemple, si une requête comme CUIVRE OU FER ne donne aucune réponse de la base de données, elle peut être élargie par le mot-clé qui est père de ceux utilisés (par exemple METAL) ou par des mots-clés associés (par exemple, les abrégés synonymes CU OU FE). Inversement si une requête donne trop de réponses, elle peut être affinée par descente dans l'arbre des mots-clés (elle peut aussi, bien sûr, être affinée par augmentation du nombre des mots-clés associés par des opérateurs ET).

Le thésaurus constitue une description du monde dont traite le système documentaire, ce qui permet les petites inférences évoquées ici. Il n'est cependant pas possible d'aller très loin dans cette voie. En effet, dès que le thésaurus dépasse la dizaine de milliers de termes, sa structuration pose de gros problèmes : on est conduit à multiplier les types de relations et les liens entre les termes s'enchevêtrent tellement que la gestion en devient impossible. Enfin, il est extrêmement

difficile de traiter les termes ayant un grand caractère de généralité, car ils impliquent des choix arbitraires graves (dans une base de données générales en sciences sociales, le mot MARXISME doit-il être fils de RELIGION, de ECONOMIE, de POLITIQUE, de PHILOSOPHIE ? Si on le met comme fils de tous, le thésaurus perd de sa valeur organisatrice. Si on fait un choix, son arbitraire est tellement grand que cela gênera profondément l'utilisateur).

A cause de ces difficultés structurelles, l'usage des thésaurus et des lexiques organisés est rare, d'autant que leurs coûts de constitution et de maintenance sont élevés.

Bien que des prototypes existent depuis le début des années 1970, ce n'est qu'au cours des années 1980 que sont apparus sur le marché des systèmes d'indexation automatique contenant des modules d'analyses plus élaborés des textes. Cette évolution est, en particulier, due à l'augmentation du nombre de systèmes documentaires contenant des textes intégraux et non plus seulement des fiches bibliographiques. Deux approches sont utilisées, le plus souvent de façon conjointe. La première est statistique. Elle consiste à faire des comptages sur les mots des textes de la base de données (en considérant grossièrement que plus un mot est utilisé dans un texte, plus il est représentatif de ce texte). Des comptages analogues sont faits sur les questions. Les résultats des comptages sont comparés et les textes considérés comme répondant à la question sont ceux dont les comptages sont les plus proches.

La seconde approche est linguistique. Des analyseurs morphologiques et syntaxiques s'efforcent d'extraire des mots-clés à partir des textes de la base de données. Ils

utilisent la théorie des grammaires en chaînes¹ et vont tenter de normaliser les mots fléchis, d'identifier les mots composés, les groupes nominaux, de retirer les mots vides. L'objectif est, évidemment, de limiter le nombre des mots-clés extraits sans pour autant que la liste des mots-clés perde sa représentativité. Les analyseurs sont également appliqués aux textes des questions. Ces deux techniques permettent donc d'utiliser comme questions des textes longs (par exemple, pour rapprocher des arrêts de cours de justice dans des bases de données juridiques, des brevets dans des bases de données sur la propriété industrielle). Elles présentent l'avantage de rendre l'indexation indépendante des fortes variations constatées pour les indexeurs humains. En contrepartie, il peut y avoir une forte perte documentaire occasionnelle, parce que l'automate, ne comprenant strictement rien au fond des textes traités, n'aura pas extrait le terme qui, justement, est le plus représentatif du texte. Ces deux techniques sont souvent utilisées de façon conjointe car les comptages seront plus pertinents s'ils sont faits après normalisation des flexions et les listes de mots-clés issues des analyseurs morpho-syntaxiques peuvent être réduites par des pondérations statistiques. A titre d'exemples de systèmes utilisant ces deux techniques, on peut citer SPIRIT et DARWIN.

On comprend que le principal problème des systèmes documentaires est que la diversité des informations qu'ils contiennent rend extrêmement difficile la réalisation d'automates qui, au-delà des traitements statistiques ou syntaxiques, réaliseraient des traitements sémantiques. Plusieurs entreprises tentent actuellement de fabriquer des

1.- Définie par Zelig Harris au début des années 1960, une grammaire en chaîne suppose que la phrase peut être décomposée en sous-chaînes éventuellement imbriquées. Une seule chaîne, dite "centrale" ne figure pas dans les autres. Elle définit le squelette de la phrase.

automates capables de construire des thésaurus ou de rédiger des résumés.

Ces travaux n'ont pour l'instant pas débouché sur des produits commercialisés.

1.5.3.4.- Lire et traduire : la traduction de textes

Si on considère que le résumé est une sorte de traduction d'une langue dans elle-même, on voit que la traduction proprement dite va rencontrer des obstacles d'un niveau encore supérieur. Dès que le texte d'origine n'est pas fortement contraint dans sa syntaxe (quelques dizaines de formes de phrases), dans son vocabulaire (quelques centaines de formes) et dans sa sémantique (pas de métaphore), il n'est pas possible aujourd'hui de le traduire automatiquement avec une bonne garantie de fiabilité. Et toujours pour les mêmes raisons : la linguistique n'a pas fourni de modèle formel suffisamment fin des langues.

Le processus de traduction d'un texte d'une langue à une autre comprend plusieurs étapes : la lecture du texte source, sa compréhension (qui, pour un humain, peut être incomplète mais jamais nulle, sauf à prendre des risques considérables d'erreurs), la préparation documentaire (recherche de dictionnaires spécialisés, de traductions dans le même domaine...), le transfert dans la langue cible et la mise en forme du texte dans cette dernière, la révision (correction des erreurs de traduction, des lourdeurs stylistiques et de l'orthographe), la saisie et la mise en forme typographique du texte final. On voit que, dans ce processus, une part importante du travail est consacrée à des tâches autres que le transfert proprement dit.

Les applications de traduction automatique ou assistée par ordinateur incluent tout ou partie de ces différentes étapes et, pour chacune d'elles, poussent plus ou moins loin l'automatisation. Comme dans tous les champs d'application des industries de la langue, une automatisation complète ne sera possible que dans des cas très particuliers. L'exemple le plus connu (et, en fait, pratiquement le seul exemple d'automatisation pratiquement complète actuellement opérationnel) est le système de traduction de la météorologie nationale au Canada. Il a été développé à l'Université de Montréal au début des années 1970 pour répondre à l'obligation de bilinguisme des textes officiels dans ce pays. Les textes sources sont les messages de prévision météo. Ils ont une syntaxe entièrement contrôlée, un vocabulaire restreint et ne contiennent aucune ambiguïté. Les textes sources sont directement accessibles sous forme électronique, ce qui élimine la double saisie en entrée. Une situation aussi favorable est évidemment rare. Elle est assez analogue à celle que rencontre un compilateur de programmes informatiques (c'est le cas idéal où le langage source est, en principe, totalement contrôlé et où tout problème de traduction ne peut résulter que d'une faute du rédacteur du texte source). Le bon fonctionnement de ce système résulte des contraintes fortes qui sont imposées au texte source. Celui-ci, en effet ne peut mettre en œuvre qu'un vocabulaire très restreint et monosémique. Sa syntaxe est entièrement contrôlée et les jeux de mots sont, évidemment, totalement exclus.

Des applications dans des situations moins contraintes ont été également développées. Parmi elles, on peut citer le système TITUS développé par l'Institut Textile de France ainsi que le système élaboré par la société Caterpillar. TITUS a d'abord été conçu pour traduire des fiches bibliographiques où les mots-

clés étaient extraits d'un thésaurus. Il a ensuite été étendu pour traduire des résumés rédigés en utilisant un vocabulaire connu du système et une syntaxe simplifiée. De même les rédacteurs des notices techniques et des manuels de formation pour les engins de travaux publics fabriqués par Caterpillar sont obligés de se limiter à un sous-ensemble de l'américain tant du point de vue du vocabulaire que de la syntaxe. Dans ces deux cas, la contrainte permet d'automatiser largement le processus de traduction. De nombreux projets de recherche continuent actuellement dans cette voie, en approfondissant l'idée qu'une interaction forte entre la machine et le rédacteur au moment de la production du texte source permettra ultérieurement de faire des traductions entièrement automatiques. Suivant les projets de recherche, l'interaction porte plutôt sur la forme du texte (la machine aide à contrôler vocabulaire et grammaire pendant la rédaction), mais peut aussi porter sur le sujet traité (la machine s'efforce de construire une représentation des connaissances contenues dans le texte).

Lorsque la contrainte sur le texte source n'existe plus, les difficultés deviennent immédiatement telles que l'on doit passer de l'idée de traduction automatique à l'idée d'assistance informatique à la traduction. Cette assistance, en premier lieu, porte sur l'environnement du traducteur : dictionnaires en ligne avec des fonctions puissantes d'enrichissement, système de traitement de texte multifenêtre permettant de visualiser simultanément plusieurs textes sur des écrans de grande dimension, logiciel de mise en page, réception et transmission téléphonique des textes, messagerie électronique et fax pour communiquer avec les clients ou l'employeur (beaucoup de traducteurs travaillent à domicile). Le système WEIDNER a été très représentatif de

cette tendance à faciliter le travail du traducteur plutôt qu'à traduire à sa place.

Cette ligne directrice a également été l'une des raisons de la création de grandes banques de données terminologiques. Elles ont été élaborées, principalement avec des financements publics, et ont été rendues accessibles à bas coût pour les professionnels de la langue (Eurodicotaum à la CEE, BTQ au Québec, etc). Elles sont cependant assez peu utilisées (temps de recherche trop long, mises à jour délicates, coût du réseau d'accès élevé), bien que des idées intéressantes aient été testées (envoi des listes de mots recherchés en télex le soir et retour des informations pendant la nuit). Des banques de terminologie développées dans des cadres moins larges semblent plus utilisées. Ainsi, la société Site à Maisons-Alfort a développé le système Phénix pour gérer la terminologie technique dont elle a besoin (Site traduit chaque année plus de 100 000 pages de documentation technique, principalement dans le domaine de l'aéronautique. Le système Phénix gère environ 250 000 termes en français, anglais, allemand, espagnol).

Depuis le début des années soixante, de nombreuses tentatives ont été faites pour réaliser également des logiciels destinés à automatiser la phase de traduction proprement dite. Trois grandes classes de techniques ont été élaborées :

- la technique du "langage pivot", qui consiste à construire une représentation du texte source qui soit indépendante de la langue d'origine. Cette représentation est ensuite utilisée pour générer le texte cible. La représentation intermédiaire peut être morpho-syntaxique ou conceptuelle. Dans les deux cas se pose le problème non résolu de la définition du langage pivot, sorte d'espéranto ou de modèle général des connaissances

humaines. Si elle était réalisable, cette hypothèse technique aurait l'avantage de permettre un multilinguisme économique puisqu'il suffirait de fabriquer un analyseur et un générateur reliant chaque langue au langage pivot.

- la technique du transfert, qui consiste à construire une représentation syntaxique du texte source dans sa langue d'origine. Chaque élément de cette représentation est traduit dans son homologue dans la langue cible, pour aboutir à une représentation syntaxique dans la langue cible, en même temps que les unités lexicales sont traduites. Un générateur transforme alors la structure syntaxique ainsi obtenue en un texte dans la langue cible. Cette hypothèse technique, qui permet également le multilinguisme, est cependant moins économique que la précédente. En effet, il faut un analyseur et un générateur dans chaque langue, plus un module de transfert pour chaque couple (orienté) de langues. C'est cette technique qui a fait l'objet du plus grand nombre de tentatives car elle semble plus accessible que la précédente et plus générale que la suivante.

- la technique du transfert strictement bilingue. En se restreignant à un seul couple orienté de langues, cette technique est assez proche de la précédente mais plus "simple" à réaliser car l'analyse du texte source peut être moins poussée et même, parfois, permet de traduire dès le niveau morphologique.

Les produits résultant de ces tentatives se répartissent en trois catégories principales :

- les petits logiciels sans prétention. Ils coûtent de 4 000 à 12 000 francs et fonctionnent sur des micro-ordinateurs en utilisant le plus souvent la troisième des techniques évoquées

ci-dessus. Très décevants dès que l'on sort de la courte phrase d'élève débutant, ils permettent néanmoins de s'amuser (si l'on en a les moyens...). Les petites machines commercialisées sous le nom de traductrices de poche se situent encore en dessous de cette catégorie puisqu'elles ne sont, pour l'instant, pas capables de faire la moindre analyse du texte source. Ce ne sont, en effet, que des dictionnaires de mots et de petites phrases pré-traduites. A cause de leur réelle portabilité et de leur spécialisation (les besoins de base d'un voyageur en pays inconnu), elles peuvent rendre cependant de réels services.

- les logiciels à prétentions professionnelles sur micro-ordinateurs haut de gamme destinés à être implantés chez l'utilisateur. On les voit régulièrement annoncés au rythme d'un ou deux par an à des prix généralement de l'ordre de 50 000 francs. Souvent issus des recherches solitaires d'un inventeur génial, ils disparaissent rapidement (qui connaît encore le logiciel de la société INTRIN, qui pourtant avait fait beaucoup de bruit lors de sa sortie en 1983-84 ?). Quelques-uns, beaucoup plus sérieux, survivent cependant avec difficulté pendant plusieurs années (ALPS, LOGOS). Leurs sociétés mères entrent même parfois en bourse (ALPS) et s'implantent dans plusieurs pays. Ces logiciels demandent un important effort d'installation à l'utilisateur qui doit lui-même constituer ses dictionnaires. En particulier, dans des applications de traduction très technique (comme la documentation d'ordinateurs), ces logiciels permettent des améliorations de productivité qui sont de l'ordre de 30 %. Leur efficacité est cependant très fortement dépendante de la persévérance et des compétences des équipes qui les utilisent.

- les systèmes en service bureau. Etant donné la difficulté et la lourdeur de la maintenance informatique et linguistique des systèmes de TAO, certaines entreprises préfèrent ne pas

vendre le logiciel mais ont mis en place des mécanismes de commercialisation en ligne ou en différé. L'utilisateur rédige, généralement sur son micro-ordinateur, le texte source. Il l'envoie ensuite au serveur de traduction qui lui retourne la traduction brute pour révision. Le principal fournisseur actuel de ce genre de service est la société Gachot à Soisy-sous-Montmorency avec le système SYSTRAN.

De très nombreux systèmes de TAO sont actuellement en phase de développement plus ou moins avancé. Les années 1980 auront été celles du redémarrage des efforts dans le domaine. Bien que l'objectif de plusieurs de ces projets soit la réalisation de systèmes à vocation générale, on reste encore aujourd'hui dans un cadre d'applications réduites à des contextes restreints.

On ne peut parler de TAO sans évoquer une application intéressante dont l'ampleur est cependant inconnue pour des raisons qui paraîtront immédiatement évidentes. Il s'agit de la traduction brute en veille technologique et en surveillance des télécommunications écrites :

Le problème de l'information scientifique et technique (dans un cadre militaire) et de la surveillance des communications internationales a été à l'origine des développements en TAO. La traduction brute issue d'un système un peu évolué, même si elle ressemble à du petit nègre, est cependant suffisante pour qu'un lecteur ne parlant absolument pas la langue source puisse évaluer grossièrement l'importance du texte qu'il regarde.

S'il est jugé important, une traduction complète en sera faite ou bien il sera lu par un pratiquant de la langue d'origine. Cette procédure de traduction brute économise du temps et des

compétences, ces deux éléments constituant un problème majeur lorsqu'il y a avalanche d'informations, comme dans le cas de la surveillance des télécommunications.

Un exemple assez typique d'application de genre est le système Tradex développé par Cap Sogéti Innovation et Cap Sesa Défense pour la Direction Générale de l'Armement française. Ce logiciel fournit des traductions brutes de télex militaires. Il fonctionne de l'anglais vers le français et il est utilisé par des spécialistes non traducteurs qui souhaitent accéder rapidement aux informations qui arrivent en permanence au central. Les textes sources (d'une demi-page à plusieurs pages) sont rédigés librement. Les phrases, généralement énonciatives, ont donc toute la variété imaginable : phases nominales, énumérations, abréviations, sigles, nombres, etc.

L'objectif est de parvenir à une compréhension d'environ 70% des phrases du télex source (cela ne signifie pas que 70% des phrases seront bien traduites, mais que 70%, bien qu'en petit nègre, seront correctement interprétées par le lecteur). La traduction est effectuée par phrase en utilisant la technique du transfert strictement bilingue. Les dictionnaires (vocabulaire courant et vocabulaire militaire) contiennent environ 5 000 mots simples et 2 000 mots composés.

Comme il n'est actuellement pas possible de traduire avec une bonne qualité les résumés, les mots-clés et les titres des ouvrages référencés dans les bases de données bibliographiques, une façon limitée de rendre partiellement multilingues des bases de données textuelles consiste à traduire préalablement les listes de mots-clés. Des listes multilingues d'équivalents des mots-clés, constituées par des traducteurs professionnels et stockées en machine,

permettent alors d'interroger, dans sa langue, une base de données rédigée dans une langue étrangère. La machine appliquera comme critère de recherche l'équivalent du terme utilisé par le demandeur. Etant donné que le document ramené par des interrogations de ce type n'est pas traduit, il faut que l'utilisateur ait une connaissance au moins passive de la langue dans laquelle est rédigée la base de données qu'il interroge. Dans cette méthode, la machine ne réalise absolument aucune traduction.

On peut évoquer, pour terminer, une fonction qui serait mise en œuvre, par exemple, dans des communications téléphoniques internationales : l'interprétation automatique simultanée.

Cette fonction, qui est aussi celle qui serait la plus utile à des scientifiques fonctionnant dans une inter-relation nécessairement planétaire, regroupe toutes les difficultés évoquées ci-dessus, depuis la machine à écrire à entrée vocale, jusqu'aux systèmes de TAO. Des projets de recherche à moyen terme incluant cet objectif avaient défrayé la chronique entre 1985 et 1987 (le projet "cinquième génération" au Japon et le projet "Alvey" au Royaume-Uni). Depuis, les échéances ont été largement repoussées ou les projets abandonnés. Toutefois, on ne peut totalement retirer du paysage des industries de la langue un projet comme celui du téléphone traducteur. Nécessitant la réunion de technologies très variées, il peut servir d'échelle de mesure des progrès. De ce point de vue, par exemple, on peut considérer que l'arrivée des réseaux téléphoniques numériques à bande étroite, puis à large bande, constituent une avancée importante, puisqu'ils permettent de transmettre une voix en haute fidélité, ce qui constitue un préalable nécessaire à tout système performant de reconnaissance de la parole.

1.5.3.5.- Ecouter, parler, lire et corriger : l'assistance à l'enseignement

L'enseignement assisté par ordinateur (EAO), tout comme la traduction automatique, est un vieux rêve dont on n'a pas fini de parler. Lorsque les ordinateurs sont devenus suffisamment accessibles pour que des pédagogues puissent s'en servir, différentes applications ont commencé à être développées. Parmi elles, les applications d'enseignement des langues. Parce que cela semblait facile, on a commencé par des conjugueurs de verbes. Il est rapidement apparu aux expérimentateurs que la conjugaison des verbes dans certaines langues (le français, au hasard...) n'était pas si simple qu'elle en avait l'air dans les livres de grammaire. Et, dans chaque langue, se présentaient des difficultés tellement considérables que, bien vite, il est devenu urgent d'attendre en se contentant de logiciels très simples. Alors, les pédagogues ont essayé de rédiger des exercices "à trou"¹, des QCM². Ils ont cependant constaté que, même dans ces procédés d'enseignement pour le moins rudimentaires, de nombreux problèmes linguistiques et d'ergonomie étaient rencontrés. Ils se sont également attaqués aux questionnaires ouverts³. Ce fut l'hécatombe. Pour parvenir à leurs fins, nos pédagogues se sont parfois transformés en programmeurs experts et en linguistes redoutables, mais le plus souvent ils ont abandonné en cours de route.

Tout au long de ces années d'exploration, des produits faisant appel au traitement automatique des langues naturelles ont

1.- Il faut compléter correctement des espaces vides dans des phrases.

2.- Questionnaire à Choix Multiple : il faut choisir la bonne réponse à une question dans une liste pré-établie.

3.- La machine pose une question à laquelle l'utilisateur répond comme il veut. La machine essaye alors d'analyser la réponse pour vérifier sa pertinence.

cependant été réalisés. Ils se répartissent en trois grande catégories :

- les applications dites fermées. Des exemples typiques sont le conjugueur de verbes, l'automate de déclinaisons latines, les exercices à trou. Ces applications permettent de faire un exercice unique et généralement très restreint en suivant un cheminement pédagogique totalement contrôlé. Les échanges entre la machine et l'apprenant sont étroitement limités pour éviter que la machine soit mise en situation de défaillance linguistique. Des recueils d'exercices de ce type pour les écoliers, les lycéens (et leurs parents...) sont commercialisés à relativement bas prix. Ils ne peuvent en aucune sorte remplacer l'enseignant mais, par leur côté souvent ludique, ils viennent varier les exercices. Ces logiciels n'aident pas seulement à apprendre les langues mais aussi le calcul, les mathématiques, la géographie. Ils ne font pas appel à des traitements linguistiques dépassant le niveau morphologique. Dans cette catégorie d'applications fermées, on peut également citer les logiciels d'aide à l'apprentissage du clavier et de la frappe à la machine qui font des chronométrages et des contrôles morphologiques sur des listes de mots (ils ne vérifient pas si vous utilisez les bons doigts sur les touches...).

- les générateurs d'applications pédagogiques aussi appelés "langages auteurs". Après avoir réalisé quelques applications fermées, en programmant en Basic, en Fortran, en LISP ou en Logo, il est logique de chercher à fabriquer un langage de programmation orienté vers la conception d'applications d'enseignement. De nombreux produits de ce type ont été réalisés. Pour pouvoir concevoir des applications d'enseignement complexes, on n'échappe cependant pas à une programmation sophistiquée, même si l'on utilise un langage

de programmation très adapté (par ailleurs difficile à concevoir). Il faut donc, de fait, être un programmeur de niveau professionnel, doublé éventuellement d'un linguiste performant. Ces générateurs d'applications ne sont donc pas utilisés par les enseignants qui étaient censés, avec leur aide, produire massivement des logiciels d'EAO. Des versions très simplifiées de ces générateurs permettent de fabriquer, sans aucune programmation, des QCM ou des exercices à trous. Après quelques années d'engouement (et de massifs soutiens publics), ces générateurs ne sont plus guère commercialisés.

- des logiciels destinés à d'autres usages qui sont utilisés à des fins éducatives. De nombreux enseignants utilisent des logiciels de traitement de texte pour concevoir des exercices de langue. De même, les logiciels d'hyper-textes connaissent un grand succès. Des environnements logiciels habituellement utilisés par les linguistes servent aux enseignants à fabriquer des listes de mots, des corpus de textes, des exemples pour préparer les cours.

A la limite des applications d'enseignement et des applications documentaires, il faut parler des dictionnaires et encyclopédies électroniques destinés au grand public. Ces produits sont apparus avec la technologie des CD-ROM. L'un des premiers produits a été le BookShelf de Microsoft qui offrait une panoplie d'aide à la rédaction allant du dictionnaire général à la liste des codes postaux américains. Encore naissant, ce marché se développe doucement mais régulièrement. Il est, entre autres, limité par les faibles connaissances actuelles en matière de traitement automatique des langues naturelles.

Des produits faisant appel à la synthèse et à la reconnaissance de la parole commencent également à apparaître. Ils étaient

jusqu'à présent limités par l'absence d'équipements performants de traitement du son sur les micro-ordinateurs grand public. Depuis septembre 1990, le constructeur Apple est le premier à offrir, en standard sur ses machines bas de gamme, les micros et les synthétiseurs indispensables. Une application typique en enseignement est la suivante : la courbe mélodique de la voix d'un professeur prononçant une phrase est représentée sur l'écran en même temps que le son est généré. L'élève doit alors faire coïncider sa propre courbe mélodique à la référence. Des prototypes existent aussi pour tester la capacité de l'élève à produire correctement des mots dans une langue étrangère. Ils sont destinés à repérer les défauts de prononciation et pourront proposer des corrections en allant jusqu'à indiquer le type de geste articulatoire à réaliser, par exemple : "placer la pointe de la langue plus en arrière pour produire tel phonème". Ces prototypes ne sont, pour l'instant, pas capables de fonctionner dans des environnements sonores ordinaires.

Les applications proposées dans le domaine des industries de la langue sont, comme nous venons de le voir, très nombreuses.

Mais la plupart d'entre elles ne fonctionnent encore que dans des contextes linguistiques et cognitifs limités. Quelles que soient les techniques auxquelles font appel les automates qui traitent le langage naturel, il est prématuré de penser que l'on peut reporter totalement sur la machine la tâche de traduire, résumer, corriger les textes. Chaque année amène cependant des idées d'utilisation dans de nouvelles niches, permettant ainsi d'accumuler des expériences, de financer de nouvelles recherches qui, peut-être, permettront dans un futur aujourd'hui encore non fixé de fabriquer des machines aux compétences plus générales. Les progrès accomplis depuis le début des années soixante sont impressionnants et très

caractéristiques des phénomènes d'innovation : l'avance est multiforme, irrégulière ; il y a une forte interdépendance entre les différents champs de recherche et d'application¹ ; les résultats obtenus résultent moins d'inventions individuelles géniales que d'un lent processus de maturation à l'échelle de l'ensemble de la communauté (ce qui ne retire rien au mérite de ceux qui savent voir, comprendre et expliciter avant tous les autres) ; après plusieurs décennies de glose théorique sur la langue, les réalisations concrètes de machines et leur usage dans l'ensemble de la société constituent un moteur d'une étonnante puissance.

Une chose semble actuellement certaine. C'est que nous en sommes aux tous premiers pas sur le long chemin de la compréhension et de la reproduction des mécanismes de langage chez l'homme. Et il est bien possible que nous soyons, dans le futur, amenés à faire des révolutions coperniciennes. Car, en effet, une grande part des progrès actuels ainsi que des applications opérationnelles résultent de l'utilisation d'ordinateurs qui fonctionnent suivant des principes dont rien ne dit qu'ils permettent de modéliser des compétences langagières aussi sophistiquées que les compétences humaines. Il est possible que les avancées de la biochimie et des neurosciences débouchent sur des conceptions des fonctions langagières très différentes des idées d'aujourd'hui et, peut-être, permettent de fabriquer des automates n'ayant rien de commun avec les actuels assemblages de métaux, de

1.- L'interdépendance est externe au domaine du traitement automatique des langues naturelles (la lecture automatique n'est pas possible sans caméra CCD, les inventaires terminologiques sans mémoires à grandes capacités, la vérification orthographique sans processeurs très rapides...), mais aussi interne (l'indexation automatique, la lecture optique, la machine à dicter obtiennent des performances considérablement améliorées par l'utilisation combinée de différents savoir-faire sur les langues).

silicium et de plastique. Cependant, chaque année amenant de nouveaux progrès dans la voie actuelle, celle-ci bénéficie d'une dynamique qui, pour prendre une expression de l'éthologie, l'entraîne à continuer de se développer tant qu'elle n'a pas saturé son environnement. C'est-à-dire jusqu'à ce que le but recherché soit atteint, ou jusqu'à ce qu'une technologie supérieure la concurrence au point de la supplanter. Ce n'est pas encore le cas.

Deuxième partie

La protection juridique des produits issus des industries de la langue

Introduction

Les investissements nécessaires à la réalisation de produits du secteur des industries de la langue ne peuvent être consentis dans un environnement juridique qui n'apporterait pas un niveau de sécurité satisfaisant.

Quelle réponse appropriée le droit peut-il donner aujourd'hui aux besoins de protection des industriels et plus particulièrement à ceux qui s'expriment dans le champ de l'innovation technologique ?

Pour réaliser les fonctions de parler, écrire, traduire, les automates utilisent des fonds lexicographiques et terminologiques ainsi que diverses formes de descriptions des langues écrites et parlées. L'activité liée à la terminologie et à la lexicographie s'appuyant, pour une part importante, sur la "création de l'esprit", il est naturel qu'une protection juridique soit recherchée en premier lieu dans le régime des œuvres de l'esprit organisé par la loi du 11 mars 1957.

Mais ces ressources linguistiques sont, de plus en plus souvent, étroitement incorporées à des ensembles de logiciels et constituent des objets qui du point de vue du droit (d'auteur) relèvent de régimes juridiques différents, voire même contradictoires.

La loi du 3 juillet 1985, en posant des règles propres pour les logiciels, est venue introduire un élément de distorsion au sein de la propriété littéraire et artistique. Ceci interdit a priori toute appréhension unitaire des œuvres informatiques complexes et, de fait, l'irruption de l'informatique dans

l'univers juridique oblige à reconsidérer la cohérence du droit des propriétés intellectuelles.

Dans le domaine des industries de la langue, la question qui se pose au Droit est de savoir si d'une manière ou d'une autre, il est possible de faire qu'à une unité économique (un automate de traduction, un correcteur orthographique...) corresponde une unité juridique (application de la loi de 1957 ou application de la loi de 1985).

L'étude des différents statuts législatifs envisageables revêt une grande importance car ils conditionnent la manière dont va opérer la protection des auteurs ainsi que l'étendue de cette protection.

Dans un premier temps, il sera fait un bref rappel de quelques principes essentiels de la réglementation relative au droit d'auteur ; en effet, l'attribution d'un droit privatif suppose que soient remplies les conditions posées par le droit de la propriété intellectuelle.

2.- La protection juridique des produits issus des industries de la langue

Dans cette seconde partie, nous rappellerons les grands principes qui gouvernent les textes français sur le droit d'auteur et nous nous efforcerons ensuite d'appliquer ces principes aux objets étudiés dans cette thèse, en examinant tout d'abord la composante logicielle puis les outils de description des langues.

2.1.- Les conditions de la protection par le droit d'auteur, rappel des principes

La protection par le droit d'auteur est accordée aux œuvres de l'esprit à condition qu'elles soient originales et qu'elles appartiennent au domaine littéraire et artistique. Cette protection ne s'étend qu'à la création de forme et en aucun cas aux idées, principes ou connaissances qui y sont exprimés. La qualité d'œuvre littéraire est acquise par la seule intervention du langage. Peu importe que l'œuvre soit exprimée à l'aide "de mots, de nombres, de symboles ou d'indices verbaux ou numériques".¹

Œuvres littéraires, les logiciels² et les dictionnaires sont aussi des œuvres de nature scientifique.³ Il en résulte, comme nous allons le voir, des difficultés d'application des conditions

1.- Définition donnée par le paragraphe 101 du *Copyright Act* des Etats Unis d'Amérique (du 19 octobre 1976).

2.- Voir la Directive européenne du 14 mai 1991 "concernant la protection juridique des programmes d'ordinateur". Selon l'article premier "les Etats membres protègent les programmes d'ordinateur par le droit d'auteur, en tant qu'œuvres littéraires".

3.- Les œuvres scientifiques sont celles dont le contenu a trait à la connaissance, qu'il s'agisse de faits, d'hypothèses, ou de constatations.

de la protection, le droit d'auteur n'étant, pour ces objets comme pour toute création intellectuelle, qu'une potentialité.

2.1.1.- La création de forme, objet du droit d'auteur

2.1.1.1.- La non-protection des idées par le droit d'auteur

Il est traditionnellement admis en droit français que les idées¹ appartiennent au domaine public et ne peuvent faire l'objet d'une appropriation juridique. Ce principe s'applique à toutes les propriétés incorporelles, qu'elles portent sur une œuvre (loi du 11 mars 1957), un dessin, un modèle (loi du 14 juillet 1909) ou une invention (loi du 2 janvier 1968).²

La propriété littéraire et artistique laisse hors de son champ d'application les idées qui, banales ou originales, demeurent "de libre parcours" et ne s'attache qu'à la forme dans laquelle ces idées se composent et s'expriment.

Les arguments qui conduisent à exclure les idées du champ de la protection du droit d'auteur sont multiples. Dans le domaine des sciences, les raisons les plus couramment invoquées sont le caractère objectif des idées ou connaissances et l'intérêt de la collectivité.

Le caractère "objectif" d'une formule chimique est aisément identifiable. Il est dû à la conjonction de deux facteurs d'une égale importance.

1.- La notion d'idée englobe le concept, la découverte, la méthode, la théorie, le procédé, la connaissance, le mode opératoire, la procédure, le savoir-faire, etc.
2.- Ainsi, ne constituent pas des inventions industrielles "les principes, découvertes et conceptions théoriques ou purement scientifiques." Article 2 de la loi du 2 janvier 1968.

- d'une part, le langage symbolique utilisé pour décrire l'idée est "objectif" (formulation abstraite), ce qui limite considérablement le nombre des formes symboliques que peut prendre l'expression et les rend toutes "équivalentes" entre elles. L'application d'un automate à l'une quelconque des formulations permet de la transformer en une autre.

- d'autre part, le phénomène décrit par cette formule fait partie de la Nature, que prétend décrire la science depuis qu'elle est née à Millet au sixième siècle avant J.C.¹

Peut-on déduire de cet exemple que toutes les idées ou connaissances scientifiques sont impersonnelles et sont "l'expression d'une vérité par essence objective" ?²

Dans le domaine scientifique, il est habituel d'établir une distinction entre les sciences de la nature et les sciences de l'homme et de la société. Dans les sciences de la nature, un principe fondamental de la connaissance scientifique est que le phénomène décrit existe en dehors de l'observateur et en dehors du système formel utilisé pour sa description. Le chercheur n'est pas un inventeur, c'est un découvreur au sens étymologique du mot, c'est à dire qu'il allume la lumière et braque le projecteur sur un fait qui, cependant, existait déjà dans l'ombre. Il y a donc une différence fondamentale entre les découvreurs (qui exhibent une propriété naturelle aux yeux des autres hommes) et les inventeurs, les créateurs (qui créent des objets non naturels : machines et procédés, tableaux et

1.- voir VERNANT Jean-Pierre, *Les origines de la pensée grecque*, Paris, Edition Maspéro, 1990. et Maury Jean-Pierre, *Comment la terre devint ronde*, Paris, Editions Gallimard, Col. La Découverte, 1989.

2.- CHERPILLOD Yvan, *L'objet du droit d'auteur*, Cedidac (Centre du droit de l'entreprise de l'Université de Lausanne), diffusé en France par LITEC, Lausanne, 1985.

poèmes). On reconnaît au découvreur un droit d'antériorité : c'est lui qui, le premier, a allumé la lumière. Mais on ne lui reconnaît bien évidemment aucun droit de propriété sur le phénomène lui-même. Cette position de principe est difficilement contestable dans le champ des sciences de la nature.¹ Elle a été étendue "naturellement" aux sciences de l'homme et de la société. Elle est alors beaucoup plus facile à contester parce que, contrairement aux sciences de la nature, les sciences de l'homme et de la société produisent des connaissances dont l'objectivité n'est pas démontrée (entre autres motifs, parce qu'elles ne disposent pas d'un formalisme cohérent pour exprimer leurs résultats). Il n'est donc plus possible de faire une distinction entre le fait, la lumière projetée sur le fait, l'acteur de l'illumination et l'observateur du résultat de cette illumination.

Le chercheur en sciences sociales peut être considéré comme un inventeur parce qu'il décrit certaines des différentes interprétations des faits sociaux², ces interprétations semblant potentiellement infinies. Pour un occidental qui vit dans une société où il y a des sociologues et des ethnologues, il paraît évident que la société existe en dehors de lui-même³. Mais la société existe-t-elle pour un Papou ? De même la famille ternaire (père-mère-enfant) existe d'une façon "incontestable" pour un français du vingtième siècle. Mais elle n'existe pas forcément pour un Bambara (nous aurions, bien

1.- Elle peut cependant être contestée dans le cas où le phénomène n'est pas perceptible directement par les sens humains, mais demande, pour être perçu, que soient utilisés des appareillages. Des sociologues, comme Bruno Latour et Harold Garfinkel, défendent l'hypothèse qu'il n'y a pas de découvreurs mais seulement des inventeurs, y compris dans les sciences de la nature. Voir, par exemple : LATOUR Bruno, *La découverte est une invention*, in Objectif Science, WorldMédia/Libération n°3, Paris, décembre 1991.

2.- La notion même de "fait social" fait l'objet de débat parmi les chercheurs en sciences humaines.

3.- Cette idée constituant le fondement de la Sociologie.

évidemment, tout aussi bien pu dire un Normand ou un Italien). Rien n'interdit cependant qu'un Bambara, un jour, propose aux autres membres de sa communauté une description de la famille bambara qui soit ternaire.

A l'opposé de cette première position, le chercheur en sciences sociales pourra contester qu'il est un inventeur et pourra se présenter comme un découvreur. Son argument principal est double. Il affirmera tout d'abord que les faits qu'il observe sont extérieurs et, ensuite, que cette extériorité est démontrée par la reconnaissance, l'approbation, que manifestent les lecteurs du texte découvrant sa découverte. Le chercheur-découvreur s'est contenté alors de faire passer le fait social du champ du non-dit, de l'implicite, au champ du perçu, du décrit.

Cette ambiguïté sur la position du chercheur en sciences de l'homme et de la société permet de démontrer que le critère de l'objectivité ne peut s'appliquer à toutes les idées et connaissances scientifiques. "Dès lors que l'on ne se confine plus dans l'exacte description de la réalité, mais au contraire qu'on interprète celle-ci, il arrive que l'on présente une idée nouvelle que l'on a tirée de son propre fonds".¹

Il existe en effet, dans le domaine des sciences autres que celles de la Nature, des idées spéculatives qui, même si elles ne révèlent pas toujours la "personnalité" de leur auteur, sont fortement dépendantes d'une culture, d'une idéologie, d'un courant de pensée, et possèdent un caractère subjectif (une

1.- CHERPILLOD Yvan, *L'objet du droit d'auteur*, CEDIDAC (Centre du droit de l'entreprise de l'Université de Lausanne), diffusé en France par LITEC, Lausanne, 1985, page 60.

individualité propre).¹ L'observation des grandes écoles de la linguistique, science dont l'objet est la description des langues naturelles, montre que les chercheurs adoptent souvent des points de vue divergents.

Le deuxième argument invoqué par la doctrine pour réfuter l'appropriation des idées par le droit d'auteur est celui de l'intérêt de la collectivité.

Pour que la science puisse progresser, les idées doivent pouvoir circuler librement. Cette argumentation permet à la doctrine de justifier la non-appropriation des idées spéculatives, des hypothèses et théories qui doivent pouvoir être librement reprises et discutées, que ces idées soient banales ou originales. Comme le souligne Yvan Cherpillod, "un droit exclusif, monopole absolu portant sur tous les éléments d'une œuvre, sans distinction, contredirait les nécessités du développement culturel, scientifique et technique. Le droit d'auteur doit, au contraire, s'inscrire en harmonie avec ce développement, chercher une voie médiane entre la protection absolue et la libre utilisation des biens intellectuels. Ceux-ci doivent faire l'objet d'un partage où l'auteur reçoit un monopole limité."²

Quels que soient les arguments utilisés par la doctrine pour justifier "le libre parcours" des idées, il est généralement admis que les théories scientifiques, les méthodes et les règles ne peuvent être protégées par un droit d'auteur. Sont

1.- Des chercheurs, comme Harold Garfinkel soutiennent qu'il en est de même dans les sciences de la Nature.

2.- CHERPILLOD Ivan, *L'objet du droit d'auteur*, CEDIDAC (Centre du droit de l'entreprise de l'Université de Lausanne), diffusé en France par LITEC, Lausanne, 1985.

également exclus de la protection les faits historiques, le fonds commun de la Science, mais aussi de la Nature Humaine.

2.1.1.2.- l'étendue de la protection

Sur le plan pratique, la distinction entre forme et idée devrait permettre pour une œuvre de l'esprit, de séparer les éléments protégés (la forme) de ceux qui ne le sont pas (le contenu). L'emprunt portant seulement sur une idée ou sur une méthode ne sera tout au plus considéré que comme un plagiat¹, non répréhensible juridiquement. La conséquence de ce principe est que le contenu d'une œuvre, reflet des idées, ne peut être protégé. L'emprunt (au sens de reproduction) de la forme d'expression peut être constitutif de contrefaçon.²

Mais comme le souligne Philippe Gaudrat³, "la frontière, tant chronologique qu'intellectuelle, entre l'idée et la forme protégée n'est pas toujours facile à tracer."

L'examen des décisions jurisprudentielles montre que les juges éprouvent des difficultés à identifier les éléments appropriables et les éléments appartenant au domaine public. Même si sur le plan théorique il est relativement facile de cerner ces différents éléments, le seuil d'appropriation est largement fixé au cas par cas.

1.- Le plagiat est une imitation habile de l'œuvre d'autrui, qui en reprend les idées mais en modifie la forme de telle sorte qu'une contrefaçon n'est pas constituée.

2.- le contrefaçon est la violation des droits de l'auteur (voir l'Article 426 du Code Pénal).

3.- GAUDRAT Philippe, *Le statut juridique des idées*, in Appropriation et circulation de l'information, Brises, CNRS, Novembre 1988, n°12, page 24.

A l'auteur d'une grammaire latine, qui accusait l'un de ses collègues de s'être inspiré de son ouvrage, les juges ont répondu que "les règles de grammaire latine sont, en quelque sorte, des principes scientifiques qui, depuis qu'ils ont été formulés, sont tombés dans le domaine public."¹

D'une façon générale, les ouvrages scientifiques bénéficient d'une protection plus limitée que les ouvrages de fiction. Comme le rappelle Eugène Ulmer, la jurisprudence craint d'en arriver à protéger un raisonnement, une constatation, autrement dit une création abstraite. L'exemple qui suit le montre clairement :

Après avoir rappelé que "les dispositions protectrices de la loi du 11 mars 1957 ne concernent pas les idées mais la forme originale sous laquelle elles sont données", les juges de la Cour d'appel de Paris, dans un arrêt du 22 janvier 1982², ont considéré que "les remèdes, méthodes, procédé d'investigation, l'exposé, même sous forme anecdotique, de cas-type tombent dans le domaine commun de la science." Toujours selon la Cour, "un ouvrage, même de vulgarisation scientifique, peut, sans encourir le reproche de reproduction systématique, utiliser le langage propre à la matière traitée, énumérer et décrire de manière identique à celles des ouvrages précédents, dans un domaine où toute précision a son importance, les caractéristiques, les propriétés, les produits thérapeutiques, les formules de traitement et leurs applications." Il ressort de cet attendu que les juges n'échappent pas à une certaine ambiguïté et entendent largement le "domaine commun" de la science. Ils reconnaissent ainsi aux tiers de larges possibilités de reproduire librement les ouvrages antérieurs.

1.- Tribunal de Nancy, 24 janvier 1904, Annales de la propriété industrielle, 1904-81.

2.- Paris, 22 janvier 1982, RIDA, Juillet 1982, n°113, page 164.

Mais l'exclusion d'une protection du fond ou contenu ne doit pas faire oublier que la forme d'une œuvre de l'esprit peut recevoir protection.

En droit d'auteur, "la forme joue un rôle capital, puisque c'est elle qui donne à l'œuvre son existence. La forme n'est pas qu'un simple moyen de communiquer l'œuvre ; elle est aussi ce qui rend déterminable la création de l'auteur. Elle permet donc de dessiner les contours du bien dont la protection est demandée, car c'est par elle que l'œuvre est perçue."¹

En littérature, qui est le domaine d'appartenance des objets étudiés dans cette thèse, la forme de l'œuvre est nécessairement celle de l'expression, ou forme externe². Littéralement, la forme d'expression d'un écrit s'interprétera comme "ordre des termes", c'est à dire une suite de mots aboutissant à une phrase. Certaines des composantes de la forme externe appartiennent de fait au domaine public ; ce sont notamment les mots, les types de versification, les accords et rythmes connus...

L'originalité³ apparaîtra alors dans le choix du vocabulaire et des expressions, la construction des phrases. Une œuvre scientifique pourra être protégée si "l'auteur a su expliciter sa pensée en des termes ingénieux de nature à frapper l'esprit du lecteur (...). Ce sera, par exemple, le cas si cet auteur a recouru à des néologismes ou enrichi sa démonstration

1 .- CHERPILLOD Ivan, *L'objet du droit d'auteur*, CEDIDAC (Centre du droit de l'entreprise de l'Université de Lausanne), diffusé en France par LITEC, Lausanne, 1985, page 74.

2.- La forme externe est le concept utilisé par la doctrine allemande pour désigner la forme d'expression d'une oeuvre de l'esprit.

3.- L'originalité est la condition nécessaire pour qu'une oeuvre puisse accéder à la protection.

d'illustrations qui sont de son cru".¹ Mais concevoir l'originalité d'une œuvre de langage uniquement à travers son expression est réducteur et aboutit à n'admettre qu'une protection partielle de la création. Le délit de contrefaçon ne pourra être retenu qu'en cas de reproduction servile de l'œuvre. Il suffira, comme le souligne André Lucas, de réécrire l'œuvre à l'aide d'un dictionnaire des synonymes pour que l'on puisse considérer qu'apparaît une autre œuvre, indépendante de la première et susceptible d'être protégée car originale dans son expression.

Admettre cette conception, c'est aussi écarter du bénéfice de la protection un certain nombre d'œuvres écrites, les anthologies ou les compilations, qui ne peuvent être originales dans l'expression mais que le législateur considère pourtant comme protégeables. Le critère de la composition ou forme interne² permet de combler cette lacune.

La composition consiste à aménager, organiser les idées. C'est notamment le plan de l'ouvrage, l'enchaînement des matières, le cheminement de la démonstration, le déroulement des événements, le caractère des personnages dans un roman...

"Au delà de l'expression, il s'agit de faire porter le droit d'auteur, non pas sur l'idée elle-même, mais sur le développement de cette idée, la composition."³ Un jugement du Tribunal de Grande Instance (TGI) de Paris⁴ donne une excellente définition de la composition : "L'œuvre de l'esprit est légalement protégeable du moment que les idées exprimées, nouvelles ou non, sont modelées dans une forme

1.- FRANCON André, *Revue trimestrielle de droit commercial*, 1982, 431.

2.- Terme employé par la doctrine allemande.

3.- LUCAS André, *La protection des créations industrielles abstraites*, Paris, Librairies Techniques, 1975, page 102.

4.- Tribunal de Paris, 9 mars 1970, RIDA, octobre 1970.

portant le sceau de la personnalité de l'auteur, original à ce titre et protégé dans la plénitude de cette originalité qui concerne d'une part le dynamisme de la composition ordonnatrice du mouvement, d'autre part la statique de l'expression achevée dans la forme définitive."

Cette opposition entre dynamisme de la composition et statique de l'expression montre bien que "la composition est moins un élément de l'œuvre créée qu'une phase dans le processus de création".¹

Si certaines œuvres littéraires, tels les romans, les poèmes, mais aussi les logiciels et les dictionnaires à usage humain, peuvent se révéler originaux à travers l'expression et la composition, en revanche, d'autres œuvres ne peuvent être originales qu'à travers une seule des composantes. Dans une traduction, par exemple, la forme externe de l'œuvre varie, mais l'organisation et la construction de l'œuvre, sa forme interne, demeure inchangée. Au contraire, une œuvre peut bénéficier de la protection légale alors que l'originalité n'apparaîtra que dans la composition. C'est le cas des compilations.

En règle générale, les législations sur le droit d'auteur protègent les compilations d'éléments à condition que le choix, l'ordonnancement et le mode ou le système de présentation de ceux-ci soient suffisamment originaux.

La loi française accorde une protection aux compilations de matières protégées, les anthologies ou recueils d'œuvres diverses qui, "par le choix et la disposition des matières,

1.- LUCAS André, *La protection des créations industrielles abstraites*, Paris, Librairies Techniques, 1975, page 110.

constituent des créations intellectuelles." Parmi les créations étudiées dans cette thèse, les corpus de références répondent à cette définition.

Les compilations contenant des matières non protégées (tels que les textes officiels) peuvent également bénéficier de la protection légale dès l'instant où leur composition révèle un minimum d'originalité. A ce titre, les tribunaux accordent une protection aux catalogues, guides, almanachs. Les dictionnaires à usage d'automates, qui contiennent uniquement des informations relevant du domaine public, entrent dans cette catégorie d'œuvres.

Mais pour les objets étudiés dans cette thèse, comme pour toutes les œuvres scientifiques, cette protection par le droit d'auteur n'est qu'une potentialité, la condition d'originalité constituant "le passage obligé"¹ de toute œuvre qui aspire à la protection.

1.- VIVANT Michel, Brèves réflexions sur le droit d'auteur suscité par le problème de la protection des logiciels, *informatica e diritto*, Florence, 1984, page 73.

2.1.2.- L'originalité de la forme

2.1.2.1.- L'originalité, condition nécessaire et suffisante

Pour bénéficier de la protection de la loi du 11 mars 1957, une œuvre doit être originale dans sa forme. Cette condition, non envisagée par la loi elle-même mais adoptée par la doctrine et la jurisprudence, est nécessaire, nulle œuvre ne pouvant y déroger, mais elle est également suffisante. Aucune exclusion ne peut être prononcée en raison de la valeur de l'œuvre, c'est-à-dire son mérite, ou de sa destination, en l'occurrence son caractère utilitaire ou industriel. Les dispositions de la loi française sont à ce sujet extrêmement claires et dépourvues d'ambiguïté. L'article 2 de la loi du 11 mars 1957 énonce que les œuvres de l'esprit sont protégées quels qu'en soient le genre, le mérite ou la destination. "Les juges n'ont pas à porter une appréciation d'ordre esthétique, à décerner des prix comme une Académie. (...) Les tribunaux n'ont pas davantage à faire une ventilation entre les œuvres qui ont une mission culturelle et celles qui ont une raison d'être purement utilitaire : protection est due aussi bien aux catalogues, aux prospectus publicitaires, aux opuscules destinés à expliquer le mode d'emploi des appareils ménagers qu'aux traités de philosophie ou de théologie".¹

Mais si la loi du 11 mars 1957 recommande de ne pas prendre en compte certains critères d'évaluation, comme la destination de l'œuvre, elle demeure malgré tout un texte sur la propriété littéraire et artistique. C'est pourquoi une partie de la doctrine essaie régulièrement de démontrer que les

1.- DESBOIS Henri, *Le droit d'auteur en France*, Paris, Ed. Dalloz, 1978, n°1.

œuvres de nature technique, scientifique ou utilitaire ne relèvent pas du droit sur la propriété littéraire et artistique, mais sont la mise en application d'un savoir-faire technique qui les met, par nature, hors du champ d'application du droit d'auteur.

Cette opinion fut développée pour la première fois par un juriste suisse, Aloïs Tröller, qui affirme "qu'il doit y avoir une représentation artistique ou littéraire qui fasse appel au sens de la beauté inhérent à l'homme".¹

Ce point de vue est également partagé en France par des auteurs comme le professeur Plaisant qui soutient "qu'il ne peut y avoir de droit d'auteur sans qu'il y ait originalité de caractère esthétique, c'est-à-dire littéraire ou artistique, sinon le domaine du droit d'auteur deviendrait sans limite".²

Le danger que dénonce l'ensemble de cette doctrine consiste en l'extension démesurée du champ d'application du droit d'auteur. Cette thèse a été défendue avant la promulgation de la loi du 3 juillet 1985, à propos des logiciels. Quelques jugements, au demeurant minoritaires, se prononcèrent pour l'exclusion du logiciel par le droit d'auteur. La Cour d'Appel de Paris jugea dans l'affaire Atari que "l'effort intellectuel du programmeur relève du savoir-faire et n'est pas de même nature que l'effort créatif protégé par le droit d'auteur".³

La nature scientifique et technique de certains produits (dictionnaires, logiciels, systèmes informatiques complexes...) les opposent irrémédiablement à la nature esthétique d'œuvres

1.- TRÖLLER Aloïs, *Urheberrecht und ontologie*, 50, UFITA, 1967, page 385.

2.- PLAISANT Robert, *La protection du logiciel par le droit d'auteur*, G.P. 83, 1, 348.

3.- C.A. Paris, 13e ch., 4 juin 1984.

plus classiques tels que les romans, les sculptures ou les peintures. Les premiers relèvent plutôt d'un effort intellectuel alors que les secondes sont plus proches d'un acte créatif, personnalisé.

Mais "si l'art est bien l'expression du sentiment, c'est-à-dire du jugement, à partir d'une réaction personnelle, cette remarque suffit à écarter toute considération esthétique dès lors que les notions de mérite et de destination de l'œuvre doivent être tenues pour indifférentes".¹ Non seulement la création n'a pas à présenter de caractère esthétique mais qu'elle en soit totalement dépourvue est parfaitement indifférent à sa protection. Celle-ci est assurée indépendamment d'une fin culturelle ou utilitaire de l'objet.

De fait, considérer la substance esthétique de l'œuvre comme une condition du droit d'auteur, s'ajoutant à la condition d'originalité, a toujours été démentie par la jurisprudence dominante et les textes officiels. Les tribunaux, antérieurement et postérieurement à la promulgation de la loi du 11 mars 1957, ont toujours été très clairs sur ce point en acceptant de protéger des œuvres comme les almanachs, indicateurs de chemins de fer ou autres compilations à but utilitaire. En 1982, la Cour de Cassation a affirmé, dans un attendu de principe, que "la protection légale s'étend à toute œuvre procédant d'une création intellectuelle originale, indépendamment de toute considération d'ordre esthétique ou artistique".²

Plusieurs articles de la loi du 11 mars 1957 reconnaissent expressément ou implicitement une protection aux œuvres

1.- Rapport du conseiller JONQUERES, R.D.P.I., 1986, n°4, p. 213.

2.- Cass. civ. 1e, 15 avril 1982, D. 1983, I.R. 93

scientifiques et techniques. L'énumération non exhaustive de l'article 3 inclut notamment les livres, brochures et autres écrits littéraires ou artistiques et scientifiques. La loi du 3 juillet 1985 a intégré les logiciels dans cette liste. Les ouvrages scientifiques sont également cités dans l'article 36 relatif au mode de rémunération particulier à certaines œuvres, dans le cadre des contrats d'exploitation du droit d'auteur. La Convention de Berne, source la plus importante du droit d'auteur international, et que la France a ratifiée, précise qu'elle couvre "toutes les productions du domaine littéraire, scientifique et artistique, quel qu'en soit le mode ou la forme d'expression".¹ La Convention Universelle², plus récente, également ratifiée par la France, se réfère de la façon la plus large, dès son préambule, à la protection des œuvres littéraires, scientifiques et artistiques.

Ainsi les créations à caractère scientifique, technique ou utilitaire, comme les œuvres artistiques, musicales ou de fiction, peuvent entrer dans le champ d'application du droit d'auteur, à condition d'être originales.

Mais il semble difficile d'apprécier l'originalité d'un logiciel, d'une carte géographique ou d'un dictionnaire avec les mêmes critères que pour une sculpture de Camille Claudel ou un roman de Marguerite Yourcenar.

Nous allons donc maintenant examiner les arguments qui sont invoqués par les juges pour qualifier une œuvre d'originale et lui permettre ainsi d'accéder à la protection.

1.- Convention de Berne du 9 octobre 1886.

2.- Convention de Genève, dite Convention Universelle, du 6 septembre 1952.

2.1.2.2.- La définition de l'originalité

En France, l'originalité est traditionnellement opposée à la nouveauté. L'originalité s'entend de manière subjective comme la marque de la personnalité de l'auteur.

La nouveauté, critère fondamental de la propriété industrielle (droit des brevets d'invention, des marques de service) se mesure objectivement et se définit par l'absence d'antériorité. Ainsi, selon l'article 8-2 de la loi de 1968 sur les brevets d'invention, "est nouveau ce qui n'a pas été rendu accessible au public par une description écrite ou orale, un usage ou tout autre moyen, avant le jour du dépôt de la demande de brevet".

Cette différenciation entre les critères d'appréciation (les uns objectifs et les autres subjectifs) s'explique par le fait que les législations sur les propriétés incorporelles protègent des objets à finalités différentes. "Ainsi se trouvent séparés la propriété littéraire et artistique, axée sur la personnalité des créateurs, dans un souci de préserver leurs droits moraux et patrimoniaux et la propriété industrielle qui, si elle accorde, dans leur intérêt propre, un monopole aux inventeurs (...) a pour finalité l'intérêt général, le progrès des sciences et des techniques".¹

Certes, les créations littéraires ou artistiques sont souvent nouvelles mais la nouveauté n'est pas une condition nécessaire.

Cette opposition entre originalité et nouveauté est particulière au droit français. Dans les pays anglo-saxons le

1.- JONQUERES Jean, *Droit d'auteur et droits voisins, la loi du 3 juillet 1985*, colloque de l'IRPI, Paris 21 et 22 novembre 1985, Librairies techniques, 1986, page 63

droit d'auteur (copyright) relève d'une conception plus économique et n'est pas soumis à la condition d'originalité, reflet de la personnalité. Le copyright a, depuis qu'il existe, été conçu comme un mécanisme d'incitation à l'investissement et il est beaucoup mieux adapté que le droit d'auteur français aux réalités nouvelles.

Œuvre portant la marque, l'empreinte de la personnalité de l'auteur, tels sont les termes employés le plus fréquemment par les juges français pour qualifier l'originalité. Mais la Cour de Cassation n'a jamais donné de définition générale de l'originalité. Une définition stricte, à supposer qu'il soit possible d'en fournir une, aurait enfermé les juges dans un cadre limité et aurait certainement présenté plus d'inconvénients que d'avantages. Le champ d'application du droit d'auteur est si vaste qu'il est difficile d'énoncer des caractéristiques communes à toutes les œuvres protégeables.

L'extrême variété de la nature des œuvres susceptibles d'être protégées par le droit d'auteur entraîne obligatoirement une diversité du concept d'originalité. "Ce concept est la représentation d'une notion à géométrie variable qui se colore différemment selon les genres concernés"¹, qu'il s'agisse de peinture, de sculpture, d'écrits scientifiques, de chorégraphies, ou de programmes d'ordinateur...

"En art, à la différence des sciences, le rôle de la personnalité est à lui seul capital. Personne d'autre que lui n'aurait pu créer l'œuvre de Mozart, en revanche Einstein n'était pas irremplaçable, car ses découvertes auraient été inéluctablement déduites tôt ou tard des lois scientifiques.

1.- Rapport du conseiller Jean JONQUERES, R.D.P.I., 1986, n°56.

L'originalité de leurs œuvres n'est pas d'essence identique".¹ Plus les œuvres sont techniques, plus la tendance vers la nouveauté s'accroît et revêt de l'importance. Mais la nouveauté du contenu ne permet pas de présumer l'originalité de la forme.

L'originalité apparaît, en jurisprudence, présumée de fait lorsqu'il s'agit pour les juges de la déceler dans des œuvres classiques, littéraires ou artistiques. Il est en effet relativement facile, sans avoir à porter de jugement sur le mérite de l'œuvre ni à s'intéresser à sa destination, de démontrer que de telles œuvres portent l'empreinte de la personnalité de l'auteur.

Par contre, dans le domaine scientifique et technique, des difficultés d'appréciation du critère d'originalité découlent du fait que des limites particulières sont imposées aux objets eux-mêmes.

Les cartes de géographie, les manuels de physique, les dictionnaires ne présentent généralement pas de caractéristiques suffisamment individuelles pour que l'on y décèle l'empreinte d'une personnalité. Comme le souligne Philippe Gaudrat "dès lors qu'il y a description rigoureuse, il n'y a plus de place pour l'arbitraire d'une manifestation personnelle dans la forme".²

Mais l'examen de la jurisprudence montre que, depuis toujours, le droit d'auteur protège des œuvres dites "mineures", les almanachs, catalogues, indicateurs dont l'originalité est perçue à travers "un important travail", un effort

1.- Rapport du conseiller Jean JONQUERES, R.D.P.I., 1986, n°54 et n°79.

2.- GAUDRAT Philippe, *Le statut juridique des idées*. In Appropriation et circulation de l'information, Brises, CNRS, novembre 1988, n°12, page 30.

"intellectuel". "De telles productions ont donné parfois le sentiment de constituer un corps étranger dans le droit de la propriété littéraire et artistique".¹

Le concept d'originalité est donc obligé de s'adapter à la nature des œuvres mais aussi à l'évolution technologique. L'exemple du logiciel, qui focalise depuis plusieurs années l'attention des juristes, est révélateur de toutes ces incertitudes. La loi du 3 juillet 1985 a mis fin à la polémique quant au régime juridique applicable, mais elle n'a pas mis fin aux difficultés soulevées par l'application des conditions de la protection. Les débats autour de la protection du logiciel ont permis d'amorcer une réflexion d'ensemble sur la notion d'originalité.

L'examen de l'étendue de la protection du logiciel ainsi que de son éventuelle originalité nous permettra de mieux comprendre l'évolution qui s'est peu à peu dessinée et d'examiner le sens qu'il convient de donner aujourd'hui à la notion d'originalité.

2.2.- La protection du logiciel

"Objet de protection physique (plombage, verrouillage...), le logiciel est aussi objet de protection juridique"².

Le choix du législateur a consisté à incorporer le logiciel à la liste des œuvres protégeables telles qu'elles figurent à l'article 3 de la loi du 11 mars 1957. Mais tout en le

1.- CHERPILLOD Ivan, *L'objet du droit d'auteur*, CEDIDAC (Centre du droit de l'entreprise de l'Université de Lausanne), diffusé en France par LITEC, Lausanne, 1985. page 88.

2.- LAMY - Droit de l'informatique, 1992, n°833.

considérant comme une œuvre de langage, à l'instar des œuvres littéraires, le législateur a tout de même estimé nécessaire de réserver au logiciel une place à part au sein de droit d'auteur ; en effet, ce type d'œuvre est soumis à un environnement et des contraintes économiques très différentes des œuvres de fiction, artistiques ou musicales.

Les dispositions de la loi du 5 juillet 1985¹ interviennent en tant que dispositions dérogatoires au régime de droit commun (copie non autorisée, durée de protection plus courte...). Mais tous les points sur lesquels la loi nouvelle n'est pas intervenue sont régis par la loi de 1957. Le silence de la loi nouvelle vaut adoption des règles anciennes, légales, doctrinales et jurisprudentielles. En conséquence, le logiciel, comme toutes les œuvres de l'esprit, doit répondre à deux exigences : avoir une forme et être original. Peu importent le langage employé, le support (disquette, ROM...) et la finalité (logiciel de base, d'exploitation...).

Une rapide description des éléments composant le logiciel et des différentes étapes que comporte le travail de programmation nous permettra de cerner les problèmes posés par l'application des textes sur le droit d'auteur.

2.2.1.- Éléments de définition technique d'un logiciel

Un logiciel est "un ensemble de programmes, procédés et règles et, éventuellement de la documentation relatif au fonctionnement d'un ensemble de traitement de données".²

1.- Loi n°85-660 du 3 juillet 1985, D. 1985, législation, 356.

2.- Arrêté du 22 décembre 1981 relatif "à l'enrichissement du vocabulaire de l'informatique" (JONC du 17 janvier 1982, page 624).

L'élément essentiel du logiciel est le programme, série ordonnée d'instructions destinées à être traitées par l'ordinateur.

Mais la définition donnée ci-dessus fait apparaître que le programme n'est qu'un élément du logiciel ; celui-ci inclut, en amont, les principes qui président à l'organisation du système dans son ensemble et, en aval, la documentation connexe ou associée.

L'élaboration d'un programme s'effectue à travers un processus de développement qui se traduit par une présentation successive de ce programme sous des formes de plus en plus appropriées à l'ordinateur et de moins en moins intelligibles à l'homme. A la base du programme se trouve l'algorithme, "ensemble fini de règles déterminées servant à résoudre un problème au moyen d'un nombre fini d'opérations" (définition AFNOR). La réalisation d'un algorithme correspond à la phase d'analyse fonctionnelle et organique du programme.

L'analyse fonctionnelle s'attache à déterminer et formaliser les opérations que le programme devra accomplir et la fonction qu'il devra remplir. L'analyse organique vise, quant à elle, à structurer le programme en fonction d'un environnement tenant compte des spécificités du matériel et des contraintes issues du programme d'exploitation. Préalable à la programmation proprement dite, cette première étape permet d'établir un schéma général de solution correspondant à un problème déterminé, défini auparavant dans un cahier des charges.

Une fois achevées, les analyses aboutissent à un document, les spécifications détaillées de réalisation, comprenant notamment l'architecture du programme sous forme

d'organigramme. L'organigramme est une représentation graphique utilisant des symboles.

Avec la deuxième étape commence la phase de programmation proprement dite. Elle consiste à reprendre directement les éléments de l'algorithme tels qu'ils sont exprimés dans l'organigramme en les traduisant en langage dit évolué (Fortran, Cobol, Lisp, Pascal...). Les instructions rédigées dans ces langages de programmation s'intitulent code source.

Afin de faciliter le travail du programmeur, ces langages ont été souvent conçus à partir de mots empruntés au langage naturel.¹ La composition du programme source correspond à une série d'instructions destinées soit au fonctionnement de l'ordinateur (c'est-à-dire l'exécution du programme par la machine ; c'est dans cette partie que l'on retrouve l'algorithme), soit à l'utilisateur qui introduit les données que le programme devra traiter.

Afin de pouvoir ensuite être utilisé par l'ordinateur, le programme source doit être converti en langage machine, qui est un langage binaire intitulé code objet. La caractéristique fondamentale du code objet est qu'il est destiné à être "lu" par l'ordinateur et non par l'homme. La conversion du code source en code objet relève d'un processus mécanique et non d'un travail humain. Cette traduction, dénommée compilation, est effectuée par la machine à l'aide d'un programme compilateur, adapté à un ordinateur donné et qui compile un langage donné.

1.- Ainsi, l'instruction PRINT, en BASIC, correspond à l'exécution d'une impression sur papier et utilise le mot anglais "*print*". Mais, il existe aussi des langages de programmation qui utilisent d'autres symboles que les mots du langage naturel pour représenter des instructions. C'est, par exemple, le cas d'APL, qui utilise massivement des caractères de l'alphabet cyrillique.

Cette description technique du logiciel nous permet de distinguer les éléments non protégés, car relevant du domaine des idées, des éléments protégés, car formalisés.

2.2.2.- L'étendue de la protection et l'originalité du logiciel

L'activité intellectuelle du programmeur réside essentiellement dans la faculté de choix qu'implique l'élaboration du logiciel. L'exercice de ces choix contribue à donner à un programme une forme particulière, distincte de celle élaborée par un autre programmeur pour l'exécution d'une même fonction. "Les analystes programmeurs ont à choisir, comme les traducteurs d'ouvrages, entres divers modes de présentation et d'expression (...), en particulier choix du langage évolué pour écrire le code source, choix des expressions de ce langage, choix des spécifications(...)"¹

Mais, contrairement aux traductions qui sont des œuvres dérivées d'une œuvre première et dont la composition (le plan) reste inchangée, l'originalité d'un logiciel peut apparaître à la fois dans l'expression et dans la composition.

Il est généralement admis que la protection doit être refusée à l'algorithme car il relève clairement du domaine des idées. Monsieur Bonneau, par exemple, dans son commentaire de la décision rendue par le Tribunal de grande instance de Paris le 21 septembre 1983, considère que "le concept que certains identifient à l'algorithme, c'est à dire les analyses

1 .- Paris, 4e chambre, 2 novembre 1982, G.P. 1983 - 1 - 117.

fonctionnelles et organiques, semble rester en dehors de toute protection".¹

Une partie de la doctrine refuse également une protection éventuelle à l'organigramme, "qui ne ferait que traduire un processus logique. Cela ne devrait pas suffire à établir une originalité qui relève du monde des formes".²

Monsieur Jean Jonquères, Rapporteur des arrêts de la Cour de Cassation du 7 mai 1986, s'est également prononcé en ce sens : "l'auteur d'un logiciel au stade de l'organigramme ou de l'algorithme ne peut interdire à quiconque de s'inspirer de son invention car le promoteur d'une idée ne peut s'opposer à son utilisation".³

Ces auteurs considèrent que c'est l'écriture de programmation qui donne sa première forme au logiciel et qui constituerait le premier état de l'œuvre logicielle. C'est donc seulement dans l'expression, à travers un langage informatique, que l'auteur pourra faire œuvre originale.

Mais la phase la plus créatrice de l'élaboration du logiciel se manifeste en amont de la programmation, dans l'algorithme et sa traduction schématisée l'organigramme. L'écriture du programme n'est possible que sur la base de l'analyse antérieure et peut être obtenue automatiquement.

Ainsi "l'analyse de la structure d'un logiciel devrait aboutir à la conclusion qu'il n'est pas protégeable par le droit d'auteur".⁴

1.- Tribunal de grande instance de Paris, 21 septembre 1983, GP.1984, 1, 51.

2.- LUCAS André, Jurisclasseur "*Propriété littéraire et artistique*", fascicule 303-1, n°36.

3.- Rapport du Conseiller Jean JONQUERES, Revue de droit de la propriété industrielle, 1986, n°49.

4.- Rapport du conseiller Jean JONQUERES, R.D.P.I., 1986, n°48.

Une telle démarche conduira en pratique à refuser le bénéfice de la protection du droit d'auteur à la plupart des logiciels.

Ces points de vue ne prennent pas en compte le fait qu'une œuvre de l'esprit peut être originale à travers son expression mais également sa composition, c'est-à-dire le plan établi par l'auteur.

Appliqué aux logiciels, cela signifie que la protection peut être accordée à l'organigramme dans la mesure où il est exprimé formellement et relève de la composition de l'œuvre. L'organigramme peut ainsi être comparé "au plan d'un ouvrage littéraire ou scientifique, c'est-à-dire un résumé de ces ouvrages, voire à un sommaire détaillé dont les symboles expriment d'une façon codée la signification".¹

C'est également ce qu'exprime Guy Boulaye lorsqu'il énonce que la structure (ou architecture) d'un programme peut être protégée en vertu de l'analogie entre, d'une part, l'organisation en modules (éventuellement en sous-programmes) et leur enchaînement et, d'autre part, la composition d'une œuvre littéraire.²

L'Assemblée Plénière de la Cour de Cassation, dans l'arrêt du 7 mars 1986, s'est clairement prononcée sur l'étendue de la protection. Utilisant une formule très large, elle semble avoir voulu protéger l'ensemble des éléments constitutifs du logiciel, exception faite de l'algorithme. Elle a en effet approuvé les juges du fond d'avoir "exactement retenu qu'il y a lieu de voir dans l'organigramme la composition du logiciel et

1.- LAMBERT Bernard, *Nature juridique des programmes d'ordinateur*, thèse de doctorat en droit, Université de Paris 2, 1969.

2.- BOULAYE Guy, *Eléments de droit pour informaticiens*, Editions Ellipses, Paris 1990, page 96.

dans les instructions rédigées, quelle qu'en soit la forme de fixation, son expression... Le programme d'ordinateur ne constitue pas une simple méthode et sa protection doit être examinée dans son ensemble".

Dans la récente directive du Conseil des Communautés européennes concernant "la protection juridique des programmes d'ordinateurs"¹, il est mentionné à l'article premier que le terme "programme d'ordinateur"², aux fins de la présente directive, comprend le matériel de conception préparatoire. Le matériel de conception préparatoire correspond au dossier d'analyse, c'est à dire, entre autres, à l'organigramme et aux tables de variables.

L'article premier de la Directive énonce également que "la protection prévue par la présente directive s'applique à toute forme d'expression d'un programme d'ordinateur. Les idées et principes qui sont à la base de quelque élément que ce soit d'un programme d'ordinateur, y compris ceux qui sont à la base de ses interfaces, ne sont pas protégés par le droit d'auteur en vertu de la présente directive." Est ainsi exclu de la protection tout ce qui est en amont du dossier d'analyse.

Mais quels que soient les éléments susceptibles d'être protégés par le droit d'auteur, il est évident qu'un logiciel ou

1.- Directive du Conseil des Communautés européennes du 14 mai 1991. Les Etats membres doivent avoir transposé la Directive dans leur droit interne avant le 1er janvier 1993.

2.- On remarquera que la Directive utilise le terme "programme d'ordinateur" et non celui de "logiciel". Concrètement, cela semble signifier qu'en France ou ailleurs, les programmes d'ordinateur devront, au plus tard au 1er janvier 1993, être protégés dans les termes de la Directive, mais que le législateur français, comme les autres législateurs nationaux, resteront libres de protéger telle ou telle facette de la création logicielle. Voir LAMY - Droit de l'informatique, 1992, n° 849.

un programme d'ordinateur ne peut refléter la personnalité de son auteur.

En effet, "l'informatique est une technique scientifique qui, comme telle, est soumise aux exigences d'une logique rigoureuse. Le programmeur est tenu à respecter les règles qui, si elles lui laissent parfois la possibilité de choisir entre deux ou plusieurs cheminements possibles, ne lui permettent que rarement d'innover ou de faire la preuve de son imagination créatrice".¹

Pour soutenir l'idée qu'un logiciel peut être original, il faut obligatoirement envisager l'une des deux analyses suivantes : "ou bien infléchir la notion traditionnelle d'originalité ou bien analyser la notion d'originalité comme ayant déjà un contenu différent de celui qu'on lui attribue usuellement".²

L'infléchissement de la notion traditionnelle d'originalité permet d'admettre qu'une œuvre est réputée originale dès l'instant où apparaît chez l'auteur de l'œuvre "un effort intellectuel individualisé".

Cette conception, soutenue par la doctrine allemande³ et en France par Christian Le Stanc, permet d'adapter le critère de l'originalité aux œuvres scientifiques et techniques. Le sens de l'originalité pourra donc varier en fonction du type d'œuvre examiné. Appliqué au logiciel, il en résulte "qu'on peut légitimement penser qu'à proportion de l'originalité de l'effort personnel du réalisateur du logiciel, la protection sera

1.- Rapport du conseiller Jean JONQUERES, R.D.P.I., 1986, n°53.

2.- LAMY - Droit de l'informatique, 1992, n°860.

3.- Voir l'article d'Eugène ULMER. *La protection par le droit d'auteur des œuvres scientifiques en général et des programmes d'ordinateur en particulier*, RIDA, 1972, n°79, pages 47 à 85.

augmentée, cette protection dépendant du poids de l'individualisation que l'on pourra observer dans la mise en œuvre de l'effort intellectuel".¹

Le problème de la protection du logiciel ne joue ici qu'un rôle de révélateur de l'existant. Le programmeur fait un (ne fait qu'un) effort intellectuel personnalisé. Mais c'est déjà ce que fait par exemple le traducteur, à qui pourtant un droit d'auteur est normalement reconnu sans discussion. C'est cet effort d'expression qui est la marque de la personnalité du traducteur et permet à la traduction d'échapper à la banalité du mot à mot. C'est aussi cette réduction d'originalité à la notion d'effort qui permet d'affirmer qu'une traduction, œuvre dérivée d'une autre œuvre, n'est que "relativement" originale.²

Cette conception de l'effort individualisé se retrouve chez Eugène Ulmer, auteur allemand : "Nous protégeons les fruits de l'effort intellectuel tel que cet effort s'exprime sous forme de paroles, d'images, ou de symboles et nous apportons à cette protection les seules limites qui sont nécessaires pour tenir compte des exigences de la vie intellectuelle et de la discussion scientifique (...). La protection du droit d'auteur ne saurait être accordée lorsque la matière même de l'œuvre en interdit une présentation individuelle".³

1.- LE STANC Christian, *La protection juridique des logiciels* - Cahiers du droit de l'entreprise, supp. n°24, 14/6/84, page 20.

2.- Une œuvre dérivée ou "composée" est une œuvre nouvelle à laquelle est incorporée une œuvre pré-existante sans la collaboration de l'auteur de cette dernière (article 9 de la Loi du 11 mars 1957). Cette œuvre est dite "relativement" originale par opposition à une œuvre "absolument" originale, c'est à dire à la fois nouvelle et originale.

3.- ULMER Eugène, *La protection par le droit d'auteur des œuvres scientifiques en général et des programmes d'ordinateur en particulier*, RIDA, octobre 1972, n°79, page 69.

Cette théorie permet d'attribuer un rôle important au critère du choix sans pour autant le rendre suffisant. L'effort intellectuel n'est pas à lui seul déterminant, contrairement à ce qui était admis pour les compilations à but utilitaire. En effet, les choix peuvent relever "de la seule aptitude professionnelle", sans pour autant révéler une individualité.

La seconde analyse, développée par Michel Vivant, semble beaucoup plus radicale. En effet, elle suggère de prendre en compte le critère de protection des brevets d'invention, la nouveauté, et de l'adapter aux œuvres de l'esprit, en insistant sur une acceptation plus objective du critère d'originalité. L'originalité ne serait qu'en fait et depuis toujours "la nouveauté dans l'univers des formes". Mais cette "objectivation" du critère de l'originalité devrait rester, en tout état de cause, limitée à la condition essentielle de forme.

Cette thèse a également été développée en Suisse par Yvan Cherpillod : "L'originalité doit être appréciée objectivement, car elle ne se détermine guère par rapport à l'auteur, mais par rapport aux biens actuellement ou virtuellement existants. La relation entre les notions d'individualité et de nouveauté devient dès lors évidente : l'œuvre est protégée si elle diffère de ce qui est connu - nouveauté - et de ce qui en découle directement - individualité - ; l'originalité est ainsi un degré supérieur de nouveauté."¹

Dans la conception classique de l'originalité, la personnalité de l'auteur doit se refléter dans l'œuvre pour que celle-ci puisse être protégée. La conception plus objective s'intéresse

1.- CHERPILLOD Ivan, *L'objet du droit d'auteur*, CEDIDAC (Centre du droit de l'entreprise de l'Université de Lausanne), diffusé en France par LITEC, Lausanne, 1985. page 133.

à l'œuvre elle-même, indépendamment de la personnalité de l'auteur.

L'Assemblée Plénière de la Cour de Cassation a, dans un arrêt du 7 mars 1986, clairement rejeté la notion classique d'originalité pour lui préférer celle plus objective d'"apport intellectuel".¹ C'est la première fois que la Cour de Cassation se prononce sur la notion d'originalité.

Dans un attendu, considéré par de nombreux juristes comme extrêmement important, les magistrats de la Cour de Cassation donnent une définition de l'originalité applicable aux logiciels : "attendu qu'en second lieu, qu'ayant recherché, comme ils y étaient tenus, si les logiciels élaborés par M. Pachot étaient originaux, les juges du fond ont souverainement estimé que leur auteur avait fait preuve d'un effort personnalisé et allant au-delà de la simple mise en œuvre d'une logique automatique et contraignante et que la matérialisation de cet effort résidait dans une structure individualisée ; qu'en l'état de ces constatations et énonciations la Cour d'Appel, qui a ainsi retenu que les logiciels conçus par Monsieur Pachot portaient la marque de son apport intellectuel, a légalement justifié sa décision".

Le conseiller Jonquères, rapporteur de l'arrêt Pachot, définit l'apport intellectuel comme étant "le fruit d'un effort pour se détacher de ses concurrents. L'effort intellectuel est ce qui fait la différence avec le travail routinier. C'est un apport de la personnalité et de l'intelligence".²

1.- C. Cass., Ass. Plén., 7 mars 1986, D. 1986, 405.

2.- Rapport du Conseiller Jean JONQUERES, Revue de droit de la propriété industrielle, 1986, n°88.

Cette conception de l'originalité permet de regrouper sous une même formule la thèse de "l'effort intellectuel individualisé" et celle de la "nouveau dans l'univers des formes". En effet, d'après le conseiller Jonquères, d'une part la "subjectivité infinie" de l'originalité n'est pas adaptée au "produit utilitaire et scientifique" qu'est le logiciel, d'autre part le terme d'originalité n'a pas un sens aussi restreint que celui qui lui est attribué habituellement. Les notions d'originalité et de nouveauté ne doivent pas être opposées et cantonnées dans un champ d'action qui leur est propre.

Toujours selon le Conseiller Jonquères, "l'originalité d'une œuvre de l'esprit naît de la combinaison de deux composantes : l'élément subjectif résultant de l'apport intellectuel personnel où se discerne la marque de l'auteur, variable selon les différents types d'œuvres de l'esprit, et le rapport chronologique avec l'existant d'où ressort l'élément objectif de la nouveauté."

En conséquence, les juges, pour accorder la protection à un logiciel, doivent procéder à deux constatations : "la première est celle de son caractère personnel mais elle va le plus souvent de soi. La seule marque de la personnalité n'est pas décisive mais elle est nécessaire. La seconde est celle d'un apport tant soit peu novateur de l'auteur dû à son effort intellectuel. L'objectivité de la seconde est de nature à éviter que la décision du juge ne soit déterminée en fonction d'un excès de subjectivité".¹ L'originalité n'étant pas faite que de subjectivité, l'appréciation de l'apport sera en conséquence relativement aisée.

1.- Rapport du conseiller Jean JONQUERES, R.D.P.I., 1986, n°95 et 51.

Ainsi comprise, la condition d'originalité, même si elle n'est pas remise en cause, paraît faire l'objet d'une interprétation "minimaliste", contrairement à ce qui se passe dans des pays comme l'Allemagne où les juges exigent que le logiciel présente un caractère hautement créatif. Tout laisse supposer "qu'un grand nombre de logiciels pourra à l'avenir être protégé par le droit d'auteur en France".¹

Il faut cependant ne pas oublier que la référence "plus noble" à l'empreinte de la personnalité de l'auteur était devenue au fil de la jurisprudence, "sous la plume de rédacteurs d'arrêts une notion "passe partout" susceptible d'attribuer arbitrairement le caractère d'originalité aux œuvres de l'esprit les plus variées par le jeu d'une simple affirmation"² qui n'était pas souvent démentie. Cette formule avait cependant le mérite de la simplicité.

La jurisprudence postérieure à l'arrêt Pachot montre que la confusion règne et que le critère de l'apport intellectuel n'est pas aisément applicable.

Certaines décisions continuent d'ignorer la directive de la Cour de Cassation, qui demandait aux juges du fond de vérifier cas par cas l'originalité des logiciels qui leur sont soumis. Ainsi, pour le Tribunal de commerce de Paris, les logiciels constituent des œuvres de l'esprit qui bénéficient en tant que telles de la protection de la loi du 11 mars 1957.³ La Cour d'Appel de Paris estime quant à elle qu'un logiciel répond par nature à l'exigence d'originalité en raison de l'effort

1.- BRIAT Martine, *Droit de l'informatique*, 1986, n°2, page 84.

2.- Rapport du Conseiller Jean JONQUERES, R.D.P.I., 1986, n°51.

3.- Tbl. Com. Paris, 1er octobre 1986, Expertises, oct. 86, page 254.

intellectuel que réclame sa rédaction.¹ On ne saurait dire plus clairement que la protection est acquise au genre.

D'autres décisions essaient de caractériser l'originalité des logiciels à travers l'apport intellectuel. Les juges hésitent entre deux tendances : soit apprécier l'originalité au regard de la nouveauté et de la pertinence des choix effectués par l'auteur, soit l'évaluer en fonction de l'importance économique du logiciel et de sa réputation.

Caractéristique de la première tendance est un arrêt de la Cour d'Appel de Paris, du 5 mars 1987, qui a jugé que le logiciel de la société Apple était nouveau, la société concurrente n'ayant pu démontrer "l'existence d'un logiciel identique". Les magistrats ont également noté que la société Apple avait négligé la "solution classique et inventé une solution économique (...). Cette solution porte la marque de la personnalité de son auteur".

Cet emprunt à la terminologie classique ne peut faire oublier que ce sont les qualités intrinsèques du logiciel qui sont prises en compte et justifient la protection.

Peu de temps après, le Tribunal de Grande Instance de Paris, ayant eu à se prononcer sur l'originalité d'un logiciel contrefait, a déclaré que l'originalité du logiciel Kortex "par rapport aux produits concurrents sur le marché mondial était de nature à permettre une conversion (d'un système d'exploitation à un autre) dite en "masse" offrant de multiples avantages."² Le tribunal a retenu l'existence d'une contrefaçon en constatant que les défendeurs "n'avaient pas la capacité

1.- Paris, 4e ch., 27 janvier 1987, J.C.P. (E) 1987 - I - 16607

2.- T.G.I. Paris, 8 avril 1987, Expertises avril 87, page 155.

technique et financière de réaliser dans un délai d'un an, un logiciel original".

La deuxième tendance, l'évaluation de l'apport intellectuel en fonction de l'importance économique du logiciel et de sa réputation, est particulièrement nette dans un jugement du Tribunal correctionnel de Paris du 8 juillet 1987.

Après avoir repris la définition donnée par l'Assemblée Plénière de la Cour de Cassation, les juges ont affirmé l'originalité des logiciels contrefaits : "attendu qu'il résulte des documents versés aux débats, qu'il s'agit de logiciels protégés car originaux, leurs auteurs ayant fait preuve, à l'évidence, d'un effort personnalisé allant au-delà de la simple mise en œuvre d'une logique automatique et contraignante (...) que notamment par exemple le progiciel "4ème dimension" qui est une base de données relationnelles est considéré comme le meilleur produit de sa catégorie, tant par sa puissance que par l'originalité des solutions adoptées pour ses fonctions".¹ Un jugement de valeur est clairement porté sur le produit contrefait. Ce sont, comme pour les autres décisions, les performances et qualités intrinsèques du logiciel qui sont prises en compte et qui permettent aux juges de déclarer les logiciels originaux.

Cependant, ces décisions révèlent une volonté certaine de la part des juges d'essayer de caractériser l'originalité à travers l'apport intellectuel. L'apport n'est-il pas défini par Jean Jonquères comme étant "le fruit d'un effort pour se détacher de ses concurrents" ? Le terme d'apport, contrairement à celui d'effort, met plus l'accent sur le résultat que sur le moyen d'y parvenir.

1.- JCP 89, I, 3376

Mais si les juges, à travers les performances des logiciels arrivent à caractériser "l'apport tant soit peu novateur de l'auteur dû à son effort intellectuel", ils ont tendance à oublier que seule une forme peut être protégée par le droit d'auteur et que le mérite d'une œuvre n'a pas à être pris en compte.

Comme le note fort justement le professeur Lucas, l'originalité pourrait être plus simplement définie par un retour aux sources du droit d'auteur. "La *summa divisio* n'est pas entre la nouveauté (objective) et l'originalité (subjective). Elle est entre le fond et la forme".¹ Il est banal de rappeler que le droit d'auteur protège les créations de forme alors que le droit des brevets protège un mécanisme. La loi de 1957 n'a pas été créée pour récompenser n'importe quel travail de l'esprit mais pour encourager les créations de forme.

Or, "la forme ne peut elle même être prise en compte que si elle est indépendante de la fonction".² C'est ce qui ressort d'un arrêt de la Cour d'Appel de Versailles. Selon les juges, "si la protection édictée par ces textes (loi de 1957) s'étend à toutes les créations originales de forme, quels qu'en soient le mérite esthétique ou la valeur artistique, elle suppose toutefois que l'auteur ait été animé, dans la conception de l'œuvre arguée d'imitation, du souci de donner à celle-ci une valeur nouvelle dans le domaine de l'agrément, et séparable du caractère fonctionnel de l'objet envisagé, ce caractère relevant éventuellement des dispositions de la loi n°68-1 du 2 janvier 1968 sur les brevets d'invention".³

1.- LUCAS André, Jurisclasseur propriété littéraire et artistique, Fasc. 303-1, n°25.

2.- LUCAS André, Jurisclasseur propriété littéraire et artistique, Fasc. 303-1, n°25.

3.- Versailles, 12e ch., 8 avril 1987, D. 1988, 202.

Il s'agissait en l'espèce, de la forme d'une planche à voile ; une création existante ayant été copiée, les juges devaient se prononcer sur le régime de protection applicable à cet objet. Il avait été relevé dans une expertise que la forme était strictement dictée par des considérations de performance optimale, de maniabilité, de rapidité et de stabilité, fins essentielles pour lesquelles l'objet avait été construit. C'était en conséquence le régime des brevets d'invention qui devait s'appliquer, sous réserve de nouveauté et d'activité inventive.

"Appliqué aux œuvres scientifiques et techniques, le critère conduira à refuser le bénéfice du droit d'auteur chaque fois que le langage mis en œuvre sera purement fonctionnel, c'est-à-dire que le fond dictera la forme".¹

Cette approche permet de relever le seuil d'originalité des logiciels et plus largement des œuvres scientifiques et techniques.

La Cour de Cassation a récemment statué en ce sens.² Les magistrats avaient à se prononcer sur l'originalité d'une compilation, en l'occurrence l'organigramme des principaux constructeurs automobiles, et à infirmer ou confirmer la définition de l'originalité telle qu'elle était apparue dans l'arrêt Pachot.

Dans cet arrêt la Cour de Cassation abandonne la conception traditionnelle de l'originalité, reflet de la personnalité de l'auteur. Elle abandonne aussi la conception trop laxiste de la jurisprudence ancienne qui admettait que les produits

1.- LUCAS André, Jurisclasseur propriété littéraire et artistique, Fasc. 303-1, n°25.

2.- Cass. civ. 1e, 2 mai 1989, D. 1989, I.R. 161.

utilitaires, comme les compilations, puissent être originaux à travers l'effort intellectuel. En effet la Cour de Cassation réfute les arguments de la Cour d'Appel ; pour les juges du fond "les organigrammes constituent bien une création originale en raison de l'effort de recherche qui a été nécessaire pour réunir les renseignements et de la composition nouvelle sous laquelle ils ont été présentés". L'arrêt est cassé au motif "qu'un travail de compilation d'informations n'est pas protégé en soi et qu'il appartenait donc à la Cour d'Appel de préciser en quoi le texte ou la forme graphique de la publication comportait un apport intellectuel de l'auteur caractérisant une création originale".

La Cour de Cassation "jette ainsi un peu de lumière sur ce que n'est pas l'originalité d'une compilation, ni la marque de la personnalité, ni un simple effort intellectuel, mais reste plus évasive sur ce que recouvre positivement l'exigence".¹ Elle réaffirme cependant clairement à travers cet arrêt que l'originalité repose sur une forme, quel que soit le mode d'expression de cette forme.

Peut-on parler à la lecture de cet arrêt d'une évolution générale du critère d'originalité et d'un abandon de la thèse "trop" subjective ? Des décisions récentes² montrent que, dans le cas d'œuvres non utilitaires, la notion d'empreinte de la personnalité est toujours utilisée. On peut en conclure que l'originalité est devenue une notion à "géométrie variable" et qu'il n'est plus possible d'envisager une unicité du critère.

L'introduction des logiciels dans la liste des œuvres protégeables par le droit d'auteur a pour effet d'accorder une

1.- LUCAS André, note sous Cass. civ. 1e, 2 mai 1989, JCP 90, II, 21392.
2.- Cass. civ. 1e, 8 décembre 1987, JCP 88, IV, 67.

importance pratique considérable à la notion d'originalité, notion qui, comme le souligne le Professeur Plaisant, n'était jusqu'à présent que d'une portée limitée.¹ La recherche de la marque d'un "apport intellectuel" laisse désormais aux juges non seulement le soin de fixer un seuil minimum d'originalité mais aussi de déterminer l'étendue de la protection des œuvres de l'esprit.

1.- PLAISANT Robert, Jurisprudence propriété littéraire et artistique, Fasc. 80, n°107.

2.3.- La protection des outils de description des langues

Dans ce chapitre, où nous examinerons les possibilités de protection des outils de description des langues par le droit d'auteur, nous reprendrons la distinction faite dans la première partie de cette thèse entre dictionnaires à usage humain et dictionnaires électroniques.

2.3.1.- Les dictionnaires à usage humain

2.3.1.1.- La "subjectivité" des ouvrages lexicographiques et terminologiques

a) les dictionnaires de langue

Nous allons introduire ici quelques éléments de définition des mots et concepts qui seront utilisés tout au long de cette thèse. Nous les adopterons provisoirement, tout en restant conscient qu'ils font l'objet d'un débat permanent et que nous ne pouvons prétendre ici donner des définitions d'aussi bonne qualité que celles qui seraient fournies par des lexicographes professionnels. On considérera donc ce qui suit comme des définitions ad-hoc fortement contextualisées par le sujet de cette thèse :

- Le dictionnaire de langue est un recueil de mots et notions, rangés dans un ordre convenu et destinés à apporter une information. Le dictionnaire est un texte utilitaire ; sa vraie nature n'est pas d'être "plein de mots" mais "plein de connaissances sur les mots".

- Les **mots** sont les formes graphiques (des suites de lettres) qui servent à désigner les notions.
- Les **notions** sont des constructions mentales qui servent à classer les objets du monde extérieur ou intérieur à l'aide d'une abstraction plus ou moins arbitraire. Dit d'une autre façon, la notion c'est le signifié, la réalité qui est étiquetée, représentée, désignée par le terme signifiant. Une notion peut être représentée par des mots différents (la notion de *chat* peut s'écrire : *chat*, *cat*...) dans des langues différentes mais également à l'intérieur d'une langue (*fauteuil éjectable*, *siège éjectable*). On dit alors qu'il y a synonymie entre les formes graphiques.
- Une **forme graphique** peut également être utilisée pour désigner plusieurs notions différentes : la forme graphique *grue* recouvre une notion d'oiseau, une notion de matériel de chantier. On dit alors qu'il y a polysémie pour la forme graphique.
- Un **mot** peut être formé d'un seul mot (*chat*) ou de plusieurs (*arbres à cames*), entre lesquels on trouve des séparateurs (espace, apostrophe, tiret...).
- Un **mot** est une suite de lettres sans séparateur entre elles. Mais cette simplicité apparente dans la définition cache une relative complexité qui apparaît lorsqu'il faut mettre au point des traitements automatiques.¹
- Lorsqu'une **nouvelle notion** apparaît dans une culture (par exemple une découverte scientifique, mais aussi une nouvelle

1.- Voir le mémoire de DEA de Nathalie CHIVA, *Un problème d'inversion de textes*, Université Paris 7, septembre 1986, publié par le CERIL (Centre d'études et de recherches en informatique linguistique).

façon de raisonner, un nouveau procédé technique...), on la désigne par des périphrases plus ou moins compliquées car la notion émergente n'est pas encore très précise. Au fur et à mesure que la notion se précise, un ou plusieurs mots sont utilisés pour la désigner.

Ces mots peuvent provenir d'emprunts à des langues étrangères (*marketing*), de recyclage de mots de la langue par analogie de forme, de notion (*un démon* en informatique, c'est une procédure qui intervient sans qu'on le lui demande, sans qu'on l'appelle, dès qu'une situation lui est favorable ; elle est donc malicieuse et imprévisible comme les démons des légendes), de construction à partir de racines, préfixes, suffixes de la langue ou de langues étrangères (*télé + vision = télévision* ; *R + A + D + A + R = radar*), de noms de marques (*mobylette*), etc. On dit alors qu'il y a néologie.

- Lorsque plusieurs mots sont utilisés dans les médias pour désigner une même notion en cours d'élaboration, après un certain temps, seul l'un d'entre eux survit. On dit alors qu'il y a **norme de fait**. C'est généralement à ce moment que le mot et la notion sont introduits dans les dictionnaires même si, parfois, l'introduction a eu lieu plus tôt (en particulier dans les dictionnaires spécialisés). C'est l'introduction dans les dictionnaires qui consacre la réalité du mot dans le langage courant ou de spécialité.

Dans le dictionnaire de langue, "les mots ne sont pas choisis pour coiffer ou titrer un discours thématique mais en tant qu'objets traités pour eux-mêmes".¹ C'est à leur propre description en tant que signe linguistique qu'ils introduisent. Le discours du dictionnaire est métalinguistique (discours sur

1.- REY Alain, *Encyclopédies et dictionnaires* - PUF, Paris, que sais-je ?, n°2000, page 20.

les mots) par opposition au discours thématique (linguistique) de l'encyclopédie qui utilise l'entrée comme simple étiquette d'un contenu.

A l'inverse de l'encyclopédie, qui "construit son image comme le cartographe fait sa carte, toujours incomplète, plus ou moins lisible, toujours arbitraire (...), on pourrait croire que le dictionnaire de langue (...) échappe à ce sort et "reflète" bien son objet, le lexique d'une ou de plusieurs langues. Mais cet objet n'est pas un. Ce n'est pas la langue dans son abstraction. C'est plutôt à une configuration complexe et conflictuelle d'usages de la langue"¹ que le dictionnaire renvoie. Ainsi que nous l'avons constaté dans la première partie, toute tentative de description d'une langue se heurte à cette caractéristique.

Le "discours" lexicographique d'un dictionnaire de langue est perceptible à travers son contenu sémantique, mais aussi à travers le choix des mots qui le compose.

Comme le montre Alain Rey, "l'élimination ou l'acceptation de mots tabous, de mots spéciaux, régionaux, archaïques, nouveaux..., le choix des éléments de discours-source (exemples et citations) permettent, quelle que soit l'apparence de rigueur de la description, de déceler l'idéologie dans tout ouvrage lexicographique... Un dictionnaire de la langue française publié en 1992 est alors un dictionnaire du lexique français fonctionnant activement et passivement à un niveau social donné en 1992, la "remontée dans le temps" se faisant en fonction du sentiment d'utilité sociale".²

1.- REY Alain, *Encyclopédies et dictionnaires* - PUF, Paris, que sais-je ?, n°2000, page 29.

2.- REY Alain, *Encyclopédies et dictionnaires* - PUF, Paris, que sais-je ?, n°2000, pages 29, 6 et 30.

Si le dictionnaire de langue peut être subjectif, car il existe un "discours" lexicographique, cette subjectivité peut également apparaître à travers le programme d'information.

Les dictionnaires de langue ne remplissent pas tous la même fonction. Chaque ouvrage possède des caractéristiques différentes, correspondant à des usages et des programmes d'information différents.

Le programme d'information, explicité dans les préfaces et avant-propos des ouvrages, concerne la nature et les objectifs du projet, le public destinataire et ses caractéristiques en termes de niveau socio-culturel ainsi que le positionnement du projet par rapport aux produits analogues.

Il définit également les contenus qui seront traités et les critères de sélection qui présideront à l'élaboration de ces contenus, c'est-à-dire le choix des entrées (nomenclature) et de la métalangue. La métalangue, langage employé pour décrire la langue dans un dictionnaire, est un langage artificiel¹, construit et conventionnel. C'est un code composé de mots de la langue ("qualité", "espèce"...). Ceux-ci sont pris conventionnellement avec un sens univoque afin de fonctionner comme mots métalinguistiques. La métalangue doit être compatible avec le niveau de compétence linguistique des utilisateurs et la métalangue conventionnelle pratiquée par tous les dictionnaires contemporains.

1.- Voir, dans la troisième partie de cette thèse, le chapitre consacré aux différences entre langage naturel et langage artificiel. Jacqueline Rey-Debove, dans son ouvrage (*La métalangue* - Edition Le Robert - Paris, 1978), le métalangage est défini comme étant "un langage dont l'objet est un langage, notamment une langue naturelle".

Le programme d'information décrit enfin tous les aspects concernant la macrostructure de l'ouvrage : le type de classement (alphabétique, thématique, notionnel, chronologique...), la typographie, la mise en page. En effet, le discours du dictionnaire est avant tout réalisé (structuré) par le langage mais aussi par d'autres codes qui peuvent avoir une certaine importance et apporter une touche "originale" à l'ouvrage.

C'est en fonction du programme d'information des différents dictionnaires usuels qu'on peut tenter de dresser une typologie¹.

Trois catégories apparaissent :

- les dictionnaires généraux qui ont pour objet la description lexicographique complète de l'ensemble des mots du lexique "général", c'est-à-dire non spécial et non spécialisé de la langue. La notion de vocabulaire général est liée à l'usage.
- les dictionnaires spéciaux qui ne s'intéressent qu'à la description d'un aspect particulier de la langue : tel type de relation entre les unités du lexique. Ce sont, entre autres, les dictionnaires analogiques et les dictionnaires de synonymes.

Les dictionnaires bilingues ou multilingues constituent également des dictionnaires spéciaux : ils recensent et mettent en relation des unités de traduction plus ou moins sommairement décrites d'un point de vue grammatical et syntaxique, mais jamais d'un point de vue sémantique.

1.- AMIEL Philippe, *Quand le dictionnaire s'automatise* - In : Les intelligences artificielles, Science et Avenir, n° spécial 53, page 99.

- les dictionnaires de spécialité qui se consacrent à la description d'un vocabulaire de spécialité : chimie, pétrole, informatique... Fondamentalement, ils recensent les mots du vocabulaire considéré sans se préoccuper de l'organisation des notions auxquels ces mots servent d'étiquette. Ils s'opposent ainsi en principe aux terminologies. Mais cette distinction n'est pas significative dans la plupart des ouvrages publiés sous le titre "dictionnaire de telle spécialité". Elle l'est en revanche pour les vraies terminologies techniques ou scientifiques qui sont rarement accessibles au public.

Les dictionnaires de langue doivent, quel que soit le programme d'information envisagé, fractionner leur discours selon un ordre résultant d'un classement. Le principe de ce classement varie selon les objets à classer : thèmes, disciplines (sciences, arts...), mots d'une langue ou de plusieurs langues.

Le découpage du dictionnaire est analogue à celui du catalogue ou de l'annuaire. Sa nomenclature¹ est une liste à laquelle sont associés des textes fragmentés. Mais ces fragments ne sont pas de même nature que ceux de l'anthologie. "Leur addition doit produire une structure répétitive et redondante..., homogène et non contradictoire (ce qui n'est pas le cas de l'anthologie) et couvrant l'ensemble d'un champ prédéterminé".²

Le "discours" lexicographique est transmis au moyen d'un arrangement textuel relativement stable, appelé couramment "article" et tenu à propos d'une entrée.

1.- La nomenclature est "l'ensemble des mots traités dans un dictionnaire" (définition du Dictionnaire actuel de la langue française, Editions Flammarion, Paris, 1990).

2.- REY Alain, *Encyclopédies et dictionnaires*, PUF, Paris, que sais-je ?, n°2000, page 10.

La description des mots d'un dictionnaire, sous forme d'articles, couvre la dimension grammaticale (catégorie, genre), le fonctionnement syntaxique (par exemple transitif ou intransitif pour les verbes), éventuellement l'histoire (étymologie) et bien sûr le contenu sémantique (définitions, exemples, citations).

b) les ouvrages terminologiques

- Un ouvrage terminologique est un recueil de termes rangés dans un ordre convenu et destiné à apporter une information.

La terminologie, par rapport à la lexicographie, se distingue plus par des besoins différents (normalisation, traduction) que par des oppositions théoriques.

Un ouvrage terminologique, comme tout ouvrage de référence (dictionnaire, lexique, encyclopédie mais aussi thésaurus...) "tente en général d'appréhender de manière significative un ensemble d'informations défini, à un certain niveau de communication, en lui conférant une homogénéité culturelle".¹

L'organisation d'un ensemble de description terminologique, la nomenclature, est fonction de la définition d'un projet. Décrire la terminologie d'un domaine suppose des décisions de frontières, ainsi que des options quant au degré de finesse de la description. Ainsi une terminologie technique décrira un domaine et ses sous-domaines à un certain niveau de précision et, parfois, plus succinctement des domaines adjacents. Le dessin du domaine est rarement donné a priori.

1.- REY Alain, *Encyclopédies et dictionnaires*, PUF, Paris, que sais-je ?, n°2000, page 26.

Sa structuration interne est souvent élaborée pour la description. "Le choix des entrées du dictionnaire présente lui-même des difficultés : les termes complexes correspondant à une notion très spécifique peuvent être inclus, certains peuvent être analysés en éléments combinables. Mais la maîtrise réelle du domaine suppose une finesse d'analyse notionnelle maximale car les termes complexes peuvent avoir une importance pratique considérable".¹

L'ordre des entrées est soit systématique, soit alphabétique. Dans le premier cas, en principe le meilleur, un classement connu des utilisateurs est souhaitable (par exemple la classification décimale universelle).

Si le dictionnaire traditionnel est un recueil d'articles décrivant des mots, le dictionnaire terminologique est, pour sa part, une collection de fiches décrivant des termes. La fiche terminologique, unité de base des ouvrages terminologiques, est comparable à l'article du dictionnaire lexicographique. A partir d'une forme considérée comme un terme et qui reçoit le statut "d'entrée", un certain nombre d'informations sont présentées de manière constante.

On peut distinguer trois niveaux d'information :

- le niveau morpho-syntaxique :
 - * la forme de l'entrée (mot ou expression)
 - * sa catégorie grammaticale
 - * son contexte
- le niveau sémantique :
 - * une définition assez brève mais précise et de

1.- REY Alain, *La terminologie, noms et notions*, PUF, que sais-je ?, n°1780.

nature relativement normative

- * une analyse de la notion par définition et des remarques éventuelles

- * le domaine d'emploi et éventuellement les sous-domaines

- * les synonymes (éventuellement)

- * les unités terminologiques d'autres langues correspondant aux mêmes concepts (équivalences)

- le niveau documentaire :

- * la source (texte officiel, nom de l'auteur de la fiche, date d'élaboration du traitement...)

Si une description générale des ouvrages lexicographiques et terminologiques a pu montrer que ceux-ci n'étaient pas la description d'un français "idéal", nous allons voir qu'une description un peu plus approfondie des contenus révèle également une part importante de subjectivité.

c) la nature des informations contenues dans les dictionnaires lexicographiques et terminologiques

Les informations contenues dans les dictionnaires lexicographiques et terminologiques, sous forme d'articles ou de fiches, peuvent être divisées en deux catégories : les informations à caractère objectif et les informations à caractère "intellectuel" ou subjectif.

Les informations à caractère objectif appartiennent au domaine public en raison de leur nature¹. Elles recouvrent la caractéristique grammaticale (adjectif, adverbe, nom

1.- Le fait est que "chaise" est un nom du genre féminin et que cette information, quelle que soit la forme sous laquelle elle est exprimée, n'a aucun caractère original.

masculin, féminin...), les synonymes du mot et les dérivations flexionnelles (conjugaison pour un verbe, forme plurielle pour un nom...). Ces informations seront consignées de manière identique dans tous les dictionnaires sauf s'il existe un usage admettant deux ou plusieurs possibilités grammaticales ou orthographiques. Ainsi le pluriel de grand-mère pourra s'écrire soit des grands-mères, soit des grand-mères, l'usage admettant l'une et l'autre graphie. Mais ces usages sont codifiés et figurent dans des textes officiels.¹

Les informations objectives englobent également le mot et le terme. Un mot est, par définition, insusceptible d'appropriation, sauf si le mot est une marque déposée, protégée en tant que telle.² Du point de vue du droit, la question qui se pose est de savoir si un mot, déposé en tant que marque, peut être librement repris dans la nomenclature d'un dictionnaire ou d'un lexique.

La marque est un signe matériel qui permet à une personne ou à une entreprise de distinguer ses marchandises ou ses services de ceux des tiers (exemple : la vache-qui-rit, ticket-restaurant).

Le dépôt d'une marque confère au titulaire un droit de propriété exclusif sur la marque déposée, pendant dix années renouvelables. Ce droit de propriété est considéré comme absolu et permet à celui qui en est titulaire d'intenter une action contre tous ceux qui y portent atteinte.

Les principales atteintes sont des actes d'usurpation qui recouvrent la contrefaçon, c'est-à-dire la reproduction servile

1.- Voir, par exemple, l'arrêté du 28/12/76 relatif aux tolérances grammaticales ou orthographiques (J.O. du 9/2/77).

2.- Loi n°64-1360 du 31 décembre 1964 ; voir Code de Commerce.

ou quasi-servile, et l'imitation frauduleuse, qui suppose un rapprochement avec une marque déjà née, donc un risque de confusion. En vertu de ces principes, les auteurs et les éditeurs de dictionnaires n'ont pas la possibilité d'insérer des marques déposées dans la liste des termes qu'ils entendent définir.

La reprise d'un terme qui est une marque déposée est pourtant pratique courante en matière de dictionnaire. Ainsi, la maison d'édition Larousse a introduit dans sa nomenclature les termes "ticket-restaurant" et "must" qui sont des marques déposées. En cas de contestation de la part du titulaire du droit de propriété, contestation très fréquente mais jamais prolongée au plan judiciaire, les éditeurs invoquent généralement deux principes complémentaires : la liberté d'information et l'usage dans le langage courant de la marque en question.

La validité du dépôt de marque est subordonnée, outre le respect des formalités, au respect des conditions posées par la loi du 31 décembre 1984. Sont ainsi exclues, faute d'originalité, les marques génériques ou nécessaires et les marques descriptives. L'appréciation du caractère générique est délicate. Cette difficulté résulte de l'évolution de la vie du mot, notamment le degré de banalisation du fait de son utilisation par le public, de son "succès", et partant de son intégration dans le domaine public comme référence commune d'identification.

Le moment auquel on doit se placer pour porter l'appréciation est déterminant. Est-ce au moment du dépôt ou au moment de l'usage critiqué ? Le choix de la solution est essentiel car il en résultera la reconnaissance ou l'exclusion d'une sorte de droit à la banalisation du mot, donc à la perte de sa protection éventuelle et corrélativement à la possibilité d'un libre usage.

La jurisprudence n'a pas admis la dégradation de fait d'un droit légal sauf négligence de son titulaire. Le caractère générique, donc le caractère de banalité, doit être apprécié au moment du dépôt¹.

Si l'originalité est reconnue, le terme est protégé. Cependant, cette protection ne sera maintenue que pour autant que le titulaire du droit n'aura pas été reconnu négligent dans la préservation de son droit exclusif. Le succès du terme-marque n'est donc pas la condamnation de sa protection. Mais l'absence d'action visant à empêcher une appropriation peut faire perdre de cette protection.

Nous considérerons alors, sous le bénéfice de ces remarques, qu'un mot, marque déposée, passé dans le langage courant (exemple : frigidaire, mobylette), peut être repris sans autorisation du propriétaire dès lors qu'il y a finalité informative.

Une analogie peut être faite avec le titre d'une œuvre de l'esprit qui est protégé comme l'œuvre elle-même dès lors qu'il présente un caractère original². L'usage admet qu'un titre puisse être librement reproduit à des fins documentaires (signalétiques). Dans un arrêt du 30 octobre 1987, la Cour de Cassation a jugé que "si le titre d'un journal ou d'un de ses articles est protégé comme l'œuvre elle-même, l'édition à des fins documentaires, par quelque moyen que ce soit, d'un index comportant la mention de ces titres en vue d'identifier les

1.- Paris, 15 mai 1971, R.T.D.C. 1971, page 1014. Cette appréciation est effectuée rétrospectivement s'il y a lieu : Paris, 20 avril 1961, R.T.D.C. 1962, page 667.

2.- Article 5 de la loi du 11 mars 1957.

œuvres répertoriées ne porte pas atteinte au droit exclusif d'exploitation par l'auteur".¹

Les informations subjectives recouvrent le contenu sémantique du terme ou du mot (c'est-à-dire le sens). Ce sont tout d'abord les définitions, les exemples et les citations mais aussi les domaines d'emploi et les équivalents dans une autre langue.

Suivant l'objectif, le public auquel est destiné le dictionnaire, celui-ci sera partagé en un certain nombre de domaines dont la somme représente une vision structurée du monde ou d'une partie du monde. Par exemple, le terme "tube reynolds" sera classé dans le domaine du cyclisme dans un dictionnaire des sports pour le grand public et dans le domaine des aciers spéciaux dans un dictionnaire de la métallurgie. Le choix de domaines (et donc d'une vision du monde) et le rattachement de chaque terme à l'un des domaines est donc bien de l'ordre du subjectif.

Un autre élément subjectif est le choix d'un terme équivalent dans une langue étrangère. Le terme anglais "cat" peut se traduire en français par "chat" mais tout aussi bien par "matou", "minet", "greffier".

Le problème de la traduction d'un mot est d'identifier dans chacune des langues les notions les plus voisines et de prendre comme équivalent les termes qui désignent ces deux notions voisines. Une langue étant la représentation d'une culture, il est clair qu'une notion en français n'a aucune notion absolument identique correspondante dans une autre langue,

1.- Cass, Ass. Plén., 30 octobre 1987, Lamy Droit de l'informatique 1991, n°2223.

mais seulement des notions voisines. On dit que les champs sémantiques des termes équivalents ne se recouvrent pas totalement. C'est d'ailleurs vrai également à l'intérieur d'une même langue : des termes donnés comme synonymes par les dictionnaires (*chat* et *minet* dans le Micro-Robert) ont des champs sémantiques qui ne se recouvrent pas totalement.

La distinction précédente entre informations objectives et informations subjectives nous permet de savoir plus précisément sur quels éléments l'originalité peut porter. Même en cas de reproduction partielle d'un dictionnaire (reprise des définitions ou de la nomenclature...), les droits des tiers (auteurs et éditeurs) doivent être respectés.

2.3.1.2.- L'originalité des dictionnaires à usage humain

a) L'originalité des dictionnaires lexicographiques et terminologiques¹

Contrairement aux compilations qui ne sont originales que dans la composition, le dictionnaire, œuvre de nature scientifique et technique, peut être à la fois original dans sa forme externe (expression) et sa forme interne (composition).

1.- Les dictionnaires de langue sont le fait d'éditeurs privés ; les dictionnaires terminologiques proviennent de sources multiples privées ou officielles. A une activité terminologique diversifiée correspondront des statuts juridiques diversifiés. Les listes de termes et définitions de terminologies d'entreprises du secteur privé pourront bénéficier de la protection de la loi de 1957 alors que les nomenclatures et vocabulaires définis par les Commissions ministérielles de terminologie et publiés sous forme d'arrêtés appartiendront au domaine public car textes officiels. Seule la réunion de ces éléments pourra faire l'objet d'une propriété au titre d'une structure originale. Voir, par exemple, le *Dictionnaire des néologismes officiels*, sous la direction d'Alain Fantapié et de Marcel Brulé, Franterm, Paris, diffusé par Nathan, 1984.

C'est ce qui ressort d'un jugement ancien qui déclare que "si le dictionnaire de Larousse est une œuvre de compilation plus ou moins savante, il constitue néanmoins un travail d'originalité incontestable, caractérisé par la mise en œuvre et l'ordre des matières, le choix des citations et des exemples, la rédaction des articles et la forme du style"¹. ces formules sont suffisamment larges pour pouvoir s'appliquer à tout ouvrage linguistique.

En littérature, le choix du mot, de l'expression revêt une grande importance et reflète précisément la personnalité de l'auteur ; ceci est particulièrement vrai pour la poésie où la sonorité du mot est primordiale.

En matière scientifique et technique, la pression de la réalité est extrêmement forte. En effet, sauf aux frontières de la recherche où plusieurs thèses peuvent co-exister, il n'y a généralement qu'une seule interprétation des faits, se ramenant à une seule expression (équation mathématique, formule chimique...). Si plusieurs visions de l'histoire, du droit, des sciences politiques restent compatibles avec la réalité des faits, il est rare que puisse être présentées plusieurs visions d'une formule chimique.

Cela n'exclut pas l'originalité mais celle-ci reposera sur "l'aménagement du détail", c'est-à-dire la façon de présenter la formule toujours identique, les exemples et les comparaisons utilisés, le cheminement pédagogique.

Si en matière de sciences sociales et humaines, la terminologie varie selon les écoles (et doit être souvent

1.- Trib. civ. Seine, 1er juin 1895, Ann. prop. ind., 1898, 303.

précisée en préambule aux ouvrages), il n'est pas de même en matière de terminologie scientifique et technique, du moins pour ce qui est de la terminologie courante (en excluant donc les dénominations à tendance commerciale ou les dénominations de concepts en cours d'élaboration).

En effet, la plus grande partie de la terminologie scientifique et technique fait immédiatement référence à des notions tangibles et visibles dont la matérialité va jusqu'au numéro de pièce mécanique, dont la réalité est incontestable et ne peut faire l'objet d'interprétations multiples. Dès que les définitions des termes scientifiques et techniques tentent une description qui ne soit pas du simple langage courant mais du langage technique, il y aura convergence forte avec pour conséquence un problème d'application du critère d'originalité.

La définition terminologique doit exprimer les traits pertinents du terme et refléter ceux de la notion (du concept) ; elle ne peut donc s'appliquer qu'à un terme "pur", parfaitement traduisible en toutes langues et qu'à un système terminologique parfaitement cohérent où chaque notion est construite de manière univoque et explicitable.

A l'inverse, la définition lexicographique explicite des signifiés tout en essayant de distinguer non pas des concepts et des classes de choses mais des sens et des classes d'usages (d'emploi) des signes. Les définitions des dictionnaires de langue, écrites en fonction du programme d'information, sont susceptibles de variations plus importantes que les définitions terminologiques.

Mais cette subjectivité est tempérée par le fait que dans un dictionnaire les définitions sont circulaires. "Une définition n'est, en principe, rien d'autre qu'une périphrase synonymique

faite de mots qui sont eux-mêmes régulièrement décrits à leur place dans le dictionnaire. Le dictionnaire organise ainsi un vaste circuit de relations entre les unités du lexique. Ce circuit rend compte du fait que le lexique est un ensemble auto-référent dont les éléments n'ont de sens que relativement aux autres éléments".¹

Les définitions finissent toujours par "boucler" sur elles-mêmes, caractéristique qui fait dire des langues naturelles qu'elles sont "réflexives". L'habileté des lexicographes consiste à faire en sorte que cette propriété ne se remarque pas trop en évitant les circularités trop rapides. Ainsi, sauf erreur, les définitions suivantes ne co-existeront pas dans un même dictionnaire :

Métal : matière travaillée par le métallier.

Métallier : activité professionnelle consistant à travailler le métal.

Les dictionnaires généraux adaptent leurs définitions en fonction de la catégorie d'utilisateurs visée dans le programme d'information. Ainsi, le mot *autruche*, pour reprendre l'exemple d'Alain Rey², ne sera pas défini dans un dictionnaire pédagogique destiné aux enfants³ mais situé par une hypéronymie (signifié plus général) : *l'autruche court très vite, un très grand oiseau*.

Dans un dictionnaire de langue en un volume⁴, la définition classe le mot par rapport au vocabulaire global mais évite

1.- AMIEL Philippe. *Vers les dictionnaires électroniques*. In Journées européennes de la traduction professionnelle, Paris 25-26 mars 1987, Encrages, Université Paris 8, n°17, 1987.

2.- REY Alain, *Encyclopédies et dictionnaires* -PUF, que sais-je ?, n°2000, p. 37.

3.- Le Nouveau Larousse des débutants.

4.- Le Petit Robert.

volontairement les caractéristiques phénoménales, factuelles et la description des caractères de l'objet ; en outre elle n'a pas à situer le terme dans une classification scientifique (taxinomie), ce qui serait indispensable en terminologie : *oiseau coureur de grande taille, à ailes rudimentaires*.

Un dictionnaire encyclopédique¹ de dimension analogue aura une pratique très différente : "oiseau ratite struthioniforme, le plus grand des oiseaux actuels (2,50 m de haut), incapable de voler mais très bon coureur (40 km/h) qui vit en bandes dans les savanes africaines et qu'on élève pour ses magnifiques plumes noires et blanches".

D'un éditeur à l'autre, des variations dans les définitions et les exemples peuvent également apparaître alors que les dictionnaires sont conçus pour une même catégorie de public.

Mais cette subjectivité suffit-elle à caractériser l'originalité d'une définition ? La tâche des juges se révèle délicate lorsqu'il est soutenu, dans un procès en contrefaçon, que des proximités sont fondées sur une cause objective et légitime, notamment en raison de la nature de la matière. L'existence de définitions voisines peut être justifiée par la nécessité d'employer un langage de spécialité.

Contrairement aux créations musicales, artistiques ou de fiction, les ouvrages à caractère scientifique et technique ont pour objet premier de transmettre des informations. La forme d'expression importe moins que le contenu. Cependant, comme nous l'avons constaté auparavant, "ce n'est pas le cœur de

1.- Le dictionnaire Hachette.

l'œuvre scientifique, mais les tissus qui l'entourent, que nous protégeons par le droit d'auteur".¹

Une œuvre scientifique, de par sa nature, ne se prête guère à une large protection par le droit d'auteur. "La protection conçue pour la forme et qui laisse libre le fond, principe fondamental de la propriété littéraire et artistique... rend la protection des œuvres de caractère scientifique d'autant plus fragile que la matière et l'essentiel de la composition échappent à l'originalité de l'auteur".²

Il a été ainsi jugé qu'un "ouvrage, même de vulgarisation scientifique peut, sans encourir le reproche de reproduction systématique, utiliser le langage propre à la matière traitée, énumérer et décrire de manière identique à celle des ouvrages précédents, dans un domaine où toute précision a son importance, les caractéristiques, les propriétés des produits thérapeutiques, les formules de traitement, leurs applications".³

Dans une décision du Tribunal de Grande Instance de Paris, en date du 24 novembre 1976⁴, les juges ont recherché "par la comparaison de deux ouvrages documentaires sur l'horlogerie, si des similitudes⁵ et identités pouvaient exister (...) et examiné si celles-ci pouvaient s'expliquer par une source commune ou une nécessité rédactionnelle liée au sujet traité,

1.- ULMER Eugène, *La protection par le droit d'auteur des oeuvres scientifiques en général et des programmes d'ordinateur en particulier* - RIDA 1972.

2.- VIVANT Michel, *Informatique et propriété intellectuelle*, JCP 84 - I - 3169, n°13.

3.- C.A. Paris (1^e ch.), 22 janvier 1982, RIDA 1982, 164.

4.- RTDC 1977, Propriété incorporelles, page 570.

5.- Selon la définition de Robert Plaisant, il y a contrefaçon aussi longtemps que les traits caractéristiques de l'oeuvre contrefaite se retrouvent dans l'oeuvre contrefaisante. Pour apprécier la contrefaçon, les juges examinent toujours les ressemblances entre les deux oeuvres et non les différences.

ou si elles étaient telles que l'on puisse considérer que l'auteur du second livre ne s'était pas livré aux opérations de l'esprit pour créer une œuvre originale et s'était purement et simplement approprié le travail de l'auteur de l'œuvre précédente, rendant inutile la réédition de celle-ci." Le Tribunal a rejeté le grief de contrefaçon au motif que les ressemblances indéniables entre les deux œuvres s'expliquaient par leur caractère documentaire.

Le 19 avril 1977, les juges de la Cour d'appel de Paris¹ ont estimé que l'existence de définitions similaires dans deux ouvrages médicaux était justifiée par l'obligation d'utiliser un langage approprié à la matière et que les emprunts étaient trop peu importants pour faire perdre son individualité propre à chacun des ouvrages.

Dans le cas des dictionnaires, la contrefaçon sera généralement établie lorsque les juges pourront prouver que les emprunts à un ouvrage antérieur sont massifs. Dans une décision ancienne, le Tribunal correctionnel de la Seine a considéré qu'un dictionnaire constituait une propriété lorsque la définition de chacun des mots qui le composent est accompagnée de développements qui sont l'œuvre particulière de l'auteur. Il y a contrefaçon dans le fait, de la part d'un tiers, d'en prendre une partie notable, en l'occurrence 99 sur 558 : "Si quelques-uns de ces mots ne contiennent que de simples définitions résumées en deux ou trois lignes et textuellement consacrées par l'usage, la plupart contiennent des développements relativement considérables, qui ne sauraient évidemment appartenir qu'à celui qui les a publiés le premier".²

1.- Arrêt inédit, base de donnée Juris-Data, distribuée par Juridial.

2.- Tbl. Corr. de la Seine, 16 août 1861, Ann. prop. ind., 1865, p. 14.

L'argument selon lequel l'auteur du dictionnaire contrefait avait fait lui-même de nombreux emprunts aux ouvrages antérieurs, et ne pouvait par conséquent revendiquer un droit de propriété, n'a pas été retenu par les juges.

Les lexicographes "se copient" les uns les autres et doivent pouvoir comme les historiens "utiliser le travail de leurs prédécesseurs".¹ Mais utiliser ne signifie pas reproduire et la contrefaçon ne sanctionne que la reproduction d'éléments formels de l'œuvre.

La contrefaçon est encore moins aisée à prouver lorsque les emprunts ne portent plus sur l'expression elle-même, à travers les définitions et les exemples, mais sur la composition du dictionnaire, c'est-à-dire sa structure et la sélection des données selon un programme d'information pré-établi.

La structure d'un ouvrage lexicographique ou terminologique est protégée à condition qu'elle révèle un minimum d'originalité. Par commodité de consultation, car un dictionnaire se consulte mais ne se lit pas, le classement par ordre alphabétique des entrées est presque toujours adopté. Les informations contenues dans les articles ou les fiches sont également organisées selon une structure relativement stable. Seuls peuvent varier les caractères typographiques ou l'utilisation de la couleur.

Mais ces touches "originales" ne permettent pas forcément d'accéder à la protection. La jurisprudence protégera un procédé inusité de classification. A l'inverse, dans l'hypothèse

1.- Paris 9 mars 1964, RIDA 1968, 199.

où l'auteur se contente de suivre un ordre habituel, d'adopter une organisation consacrée par l'usage professionnel, l'originalité est déniée. Dans le cas où la structure semble banale, reste l'éventualité de l'originalité dans la sélection des informations qui figurent dans le dictionnaire. L'élaboration d'un outil de description de la langue implique de nombreux choix, effectués en fonction du programme d'information : choix des domaines d'emploi, des équivalents pour un dictionnaire bilingue, de la nomenclature, des citations... Dans le cas de la nomenclature, ce n'est plus le choix d'une expression à la place d'une autre permettant de donner une tournure particulière à une définition, mais un choix quantitatif qui s'impose : la quantité de masse lexicale à retenir.

Pour les éditeurs de dictionnaires et de lexiques, la nomenclature représente un important travail de sélection. Certains éditeurs n'hésitent pas à piéger leurs listes d'entrées en y introduisant des mots inconnus ou en commettant volontairement des fautes d'orthographe afin de mieux se prémunir contre d'éventuels "pillages".

La nomenclature d'un dictionnaire est de fait plus ou moins imposée. Elle doit rendre compte du lexique de la langue, de son vocabulaire, en tenant toujours compte d'un certain nombre de principes de cohérence, au premier rang desquels figure la fréquence.

Un dictionnaire pour les jeunes comportera un vocabulaire de 2 500 mots, 20 000 mots pour les élèves du secondaire, de 30 000 à 60 000 mots pour les dictionnaires généraux en un volume et 100 000 mots pour un grand dictionnaire en plusieurs volumes.

De même que pour les nomenclatures des dictionnaires de langue, l'originalité des dictionnaires bilingues et des lexiques ne pourra apparaître que dans le choix des listes de mots.

b) L'originalité des lexiques bilingues et unilingues

L'usage éditorial intitule "dictionnaire" des objets fondamentalement hétérogènes ; c'est le cas pour les lexiques qui se distinguent des dictionnaires usuels par l'absence d'informations de nature sémantique.

Un lexique, dictionnaire succinct afférent à une science, un art..., unilingue ou bilingue, ne contiendra ni définitions, ni exemples, ni citations. Le lexicographe, traduisant une liste de mots ou de termes dans une langue donnée, aux fins de publication dans un dictionnaire ou un lexique bilingue, opère éventuellement un choix entre plusieurs équivalents et retient celui qui lui semble le plus pertinent. Auparavant, les mots du lexique auront été sélectionnés en fonction d'un programme d'information (domaine de référence, public visé).

Le choix opéré par le lexicographe, dont le résultat est une juxtaposition de mots sous forme de liste peut-il être considéré comme l'élément déterminant la qualification d'une œuvre originale ?

Une comparaison peut être faite avec les outils documentaires élaborés par les documentalistes dans la mesure où leurs formes sont extrêmement proches de celles des lexiques usuels.

Les principaux instruments utilisés dans un centre de documentation pour analyser et rechercher des documents sont

les thésaurus¹ et les lexiques². Les thésaurus et les lexiques sont des instruments de référence au même titre que les dictionnaires et les encyclopédies.

Ces outils documentaires sont, en général, élaborés à partir d'un programme d'information. Le documentaliste doit tenir compte du domaine couvert par le centre de documentation et en cerner les limites. Il doit également déterminer, avec autant de précision que possible, la nature du public ainsi que les besoins de ce public.

La définition d'un programme d'information (choix du domaine de référence et niveau d'analyse) permet ensuite au documentaliste de choisir les termes du lexique. Les méthodes utilisées sont proches de celles des lexicographes : se servir de listes existantes dans le domaine (dictionnaires, tables analytiques d'ouvrages, index...) et lire un grand nombre de textes du domaine considéré pour en extraire les mots qui paraissent les plus significatifs.

Les mots ou expressions sélectionnés doivent ensuite être transformés en termes du langage documentaire. Le résultat final prend la forme d'une liste de termes normalisés (appelés "descripteurs") dénommée soit "lexique", soit "thésaurus".

Le lexique, outil de base du documentaliste, est ensuite utilisé pour analyser le fonds documentaire.

1.- Le thésaurus est un dictionnaire qui à tout mot du langage naturel (dans un domaine donné) fait correspondre un (seul) terme du langage documentaire. Les termes du langage documentaire sont organisés par des relations analytiques (synonymie, hiérarchie, ...).

2.- Le lexique est un ensemble de descripteurs utilisés pour l'indexation. Les termes qui composent le lexique sont donnés par ordre alphabétique simple et ne sont pas reliés les uns aux autres par des relations analytiques.

L'analyse documentaire est une opération ou un ensemble d'opérations visant à représenter le contenu d'un document sous une forme différente de sa forme originelle.

Elle représente une opération intellectuelle plus ou moins compliquée selon son degré d'élaboration et de finesse et aboutit à un sous-produit ou document secondaire (par opposition au document analysé, dénommé document primaire). Les documents analysés sont écrits en langage naturel, alors que l'analyse documentaire fait appel aux langages documentaires.

Le langage documentaire, système de symboles, n'est pas "un langage autonome", au sens strict du terme ; l'expression "langage documentaire" est seulement fondée sur une opposition de fait entre le caractère généralement restreint du vocabulaire et/ou de la grammaire mis en œuvre pour construire les représentations indexées et la richesse du "langage naturel" dont le langage documentaire "est inspiré"¹.

Contrairement au langage naturel qui a une construction essentiellement linéaire, le langage documentaire "présente un caractère vertical dominant, qui vise à réduire le discours linéaire"². Les inconvénients et anomalies du langage naturel sont éliminés.

La méthode d'analyse documentaire la plus utilisée est l'indexation. L'indexation est l'expression plus ou moins condensée des caractéristiques d'un document dans les termes d'un langage propre au système considéré et fortement

1.- CATTENAT Annette, PAUL Gérard, *Etude de quelques langages documentaires*, Université Paris VIII, 1977, p. 168.

2.- CHAUMIER Jacques, *Les techniques documentaires*, PUF, Paris, que sais-je ?, n°1419, p. 47.

restreint par rapport au langage naturel. "Elle consiste essentiellement à dégager les notions les plus importantes d'un document à l'aide de termes empruntés à un lexique documentaire (et parfois reliés par des indications syntaxiques)¹.

L'indexation peut également être réalisée sans l'aide d'un langage documentaire. Les termes indexés sont alors choisis librement ou extraits du document lui-même. Ces index ou listes de mots-clés ne sont pas des langages documentaires.

L'élaboration d'un lexique ou d'un thésaurus est une opération intellectuelle qui implique de nombreux choix. Ces choix sont, par définition, subjectifs. Dans le cas de l'élaboration d'une liste de mots-clés, les choix sont évidemment beaucoup plus restreints car la quantité d'information à retenir est beaucoup moins importante. Cependant, l'indexation d'un document, élaborée à partir d'une liste de termes pré-établie ou à partir du texte primaire, nécessite un savoir-faire ainsi qu'une connaissance du domaine étudié. Ce savoir-faire peut-il aboutir à un résultat suffisamment original ?

Récemment, la Cour de Cassation² a jugé qu'un travail d'indexation par mots-clés (indexation d'articles du journal *Le Monde* et du *Monde Diplomatique*) était libre ; cela signifie qu'un documentaliste n'a pas à demander l'autorisation de l'auteur d'un texte source pour indexer le document, les mots-clés ne mettant en valeur que les idées du texte et ne reproduisant en aucun cas la forme d'expression, seule protégée par le droit d'auteur. Cette argumentation est tout à

1.- CATTENAT Annette, PAUL Gérard, *Etude de quelques langages documentaires*, Université Paris VIII, 1977, p. 168.

2.- Cass. 1^e civ., 9 novembre 1983, JCP 1984, II, 20189.

fait logique et n'est qu'une application du principe de non-protection des idées.

Mais la non reconnaissance d'une œuvre dérivée ne doit pas aboutir à la conclusion qu'il n'y a pas d'œuvre du tout. L'index possède indéniablement une forme qui lui est propre. "Il ne s'agit plus de l'œuvre elle-même... mais que l'indexation soit "autre chose", permet-il de conclure qu'elle devient elle-même une œuvre susceptible de protection ?"¹.

Plusieurs auteurs se sont prononcés en faveur d'une non-protection des index. Il semblerait que les conditions requises pour qu'une liste de mots-clés bénéficie d'un droit privatif puissent être difficilement remplies.

Pour Michel Vivant, "l'indexation paraît une œuvre trop sèche, trop étrangère à la personnalité de son auteur, pour pouvoir franchir le seuil d'originalité qui lui permettrait d'accéder à la protection, sauf à faire valoir que la jurisprudence a bien admis la protection par droit d'auteur de travaux tels que des indicateurs et que l'originalité -comme la nostalgie- n'est plus ce qu'elle était"².

Philippe Gaudrat estime que l'indexation, dont l'apport intellectuel est indéniable, "ne manipule que des mots-clés représentatifs de concepts non appropriables, sans aucune forme linguistique et dans le seul but de mettre en œuvre une méthode de classement"³.

1.- LAMY - Droit de l'informatique, 1991, n°2309.

2.- VIVANT Michel, *Informatique et propriété intellectuelle* - JCP 1984, éd. G., n°23.

3.- GAUDRAT Philippe, *Droits et obligations de l'utilisateur d'une banque de données* - In : appropriation et circulation de l'information, Brises, CNRS, novembre 1988, n°12, p. 109.

Alors que les auteurs défavorables à une protection mettent l'accent sur l'absence de forme d'expression et l'absence d'originalité, les auteurs qui se prononcent en faveur d'une protection de l'indexation basent leurs arguments sur l'existence d'un savoir-faire documentaire et l'existence de choix subjectifs.

"Certains ont voulu voir dans l'indexation une tâche banale, sans marge de liberté. C'est méconnaître la pratique. Les règles d'indexation varient... et, à l'intérieur d'un même corps de règles, les résultats ne sont pas identiques suivant les documentalistes. (...) Les index seront-ils issus des seuls mots du texte ou seront-ils enrichis des notions implicites ? (...) L'indexation sera-t-elle opérée à vocabulaire fermé à partir d'un lexique préétabli ou au contraire à vocabulaire ouvert ? (...) Le choix des règles d'indexation, l'élaboration des index nécessitent un effort intellectuel personnalisé dont l'extraction des notions implicites n'est pas le moindre."¹

Pour Pierre Catala, derrière le choix et l'agencement non arbitraires de mots-clés, se trouve sous-jacent un savoir-faire, matérialisé par des nomenclatures et arborescences qui sont le produit d'une réflexion et constituent une œuvre de l'esprit.

Si les auteurs du Lamy-Droit de l'informatique rejettent la protection d'un index, ils considèrent par contre qu'un *abstract*, outil documentaire proche du résumé "ne doit être tenu comme insusceptible de donner prise au droit d'auteur"

1.- MEHL-MIGNOT (H.), *L'informatique au regard de la propriété intellectuelle : banques de données et droit d'auteur* - In : appropriation et circulation de l'information, Brises, CNRS, novembre 1982, n°12, page 80.

car il possède une "forme littéraire, et il est lisible même s'il ne suit pas les canons de la grammaire"¹.

La technique de l'*abstract* est essentiellement utilisée dans les banques de données juridiques. L'*abstract* est "un descriptif analytique élaboré suivant une structure logique qui va du juridique au factuel et du général au particulier"². Il consiste à condenser l'information contenue dans les décisions de justice en faisant apparaître à la fois les concepts juridiquement fondamentaux, les éléments de fait essentiels et la solution dégagée.

L'élaboration de l'*abstract* nécessite de choisir les termes descripteurs pour formuler les données en respectant dans la mesure du possible le vocabulaire employé par le juge ou le législateur. Pour ce faire, l'analyste est aidé par des structures d'analyse. L'*abstract* doit "coller" précisément au texte initial. Il est presque toujours assorti d'un résumé.

A la différence de l'indexation pure, la technique de l'*abstract* ne se contente pas de juxtaposer les mots-clés retenus mais leur ajoute un certain nombre de mots nuls, sans lesquels il serait inintelligible. D'autre part, l'analyste n'est pas tenu par un vocabulaire fermé, à l'inverse de la technique d'indexation classique, et peut donc être le plus explicite possible tout en respectant l'esprit du texte analysé. Ainsi un *abstract* donnera : "vente d'immeuble à construire, vente à l'état futur d'achèvement, infiltrations d'air par la porte fenêtre, mauvaise fixation d'une grande vitre..." alors qu'un index donnera : "convention, validité, prestation, licéité, appropriation, interdit."

1.- LAMY - Droit de l'informatique 1989.

2.- LAMY - Droit de l'informatique 1989.

Les différences constatées ci-dessus permettent-elles dans un cas d'accorder une protection et dans l'autre la rejeter ? C'est jouer sur les mots.

La question qu'il faut poser est de savoir si le critère du choix peut être un critère suffisant pour caractériser l'originalité d'une création intellectuelle.

Le critère du choix est mentionné dans la loi du 11 mars 1957. L'article 4 de la loi du 11 mars 1957 investit du droit d'auteur les anthologies ou recueils d'œuvres diverses, qui peuvent être originaux par la composition. Cependant, cet article exige une double condition d'originalité : le choix et la disposition des matières. La conjonction "et" est tout à fait voulue par le législateur.

Cette formule restrictive n'avait pas été prévue dans le projet de loi qui employait la conjonction "ou". Mais le rapporteur de la Commission de lois avait fait observer que protéger une œuvre à travers le seul choix des matières aurait été excessif ; il ne peut y avoir création intellectuelle que si la disposition des matières s'ajoute à leur choix.¹ Une anthologie qui reprendrait des passages des œuvres d'un auteur, dans l'ordre chronologique où celui-ci les a publiées, ne ferait pas une œuvre originale.

La doctrine a d'une façon générale toujours jugée abusive l'exigence de cumul des conditions. Pour Henri Desbois, par exemple, la marque de la personnalité apparaît déjà dans le seul choix des matières ; la sélection est en elle-même un

1.- Doc. Parl., Ass. Nat., annexe au P.V. de séance du 16 février 1956, 553, page 3.

critère d'originalité. "Les juges commettraient-ils le péché d'hérésie s'ils faisaient abstraction et de la substitution d'un mot à un autre et du commentaire qui en a été donné par une voix autorisée dans l'enceinte du Parlement ? L'absolution, en tout cas, pourrait leur être donnée sans scandale car à l'interprétation littérale le raisonnement d'analogie fait contrepoids. Le compilateur révèle ses préférences, ses conceptions, son idéologie, non seulement lorsqu'il groupe, abstraction faite de l'ordre chronologique, les extraits d'œuvres diverses en fonction d'une perspective personnalisée, mais aussi lorsqu'il se contente de les présenter en tenant compte des dates respectives de publication : la sélection à laquelle il procède exprime sa personnalité ; un concurrent n'aurait pas nécessairement adopté le même choix.

La même observation est valable pour la sélection d'extraits d'un seul et même ouvrage. C'est l'esprit plutôt que la lettre des observations du rapporteur qu'auront à retenir les tribunaux".¹

On peut noter que les dispositions de l'article 2 § 5 de la convention de Berne n'exigent pas le cumul des deux conditions : "les recueils d'œuvres littéraires ou artistiques tels que les encyclopédies ou anthologies qui, par le choix ou la disposition des matières, constituent des créations intellectuelles sont protégées comme telles".

Reconnaître la protection potentielle à travers le critère du choix, "c'est largement ouvrir la voie de la protection aux œuvres de sélection. (...) Toute anthologie qui, par hypothèse, est une sélection et qui offre un visage nouveau par rapport

1.- DESBOIS Henri, *Le droit d'auteur en France* - Dalloz 1978, n°30.

aux anthologies, aux sélections antérieures, a vocation à se voir protégée".¹

Mais le cas de l'anthologie ou du recueil d'œuvres est un peu particulier car l'auteur de l'anthologie, œuvre dérivée d'une œuvre première elle-même protégée, a la possibilité, à travers son choix, de révéler sa personnalité, ses goûts intellectuels et met surtout en valeur des éléments déjà protégés par le droit d'auteur.

Une analogie peut être faite avec le traducteur littéraire, qui procède à un choix du vocabulaire et des tournures de phrases les plus appropriés. Si le traducteur a la possibilité de révéler sa personnalité et de faire œuvre originale, c'est uniquement parce qu'il traduit une œuvre elle-même originale. "Ce que la loi protège dans l'œuvre dérivée du traducteur, c'est moins le choix portant sur les mots, bien qu'il soit loin d'être indifférent, que l'expression qui doit restituer fidèlement l'atmosphère de l'œuvre originale. C'est cet effort d'expression qui est la marque de la personnalité du traducteur et qui permet d'échapper à la banalité du mot à mot. Sans cette condition, la traduction ne saurait pas plus être protégée que celle d'une banale notice ou d'un mode d'emploi".²

Dans le cas des compilations d'éléments appartenant au domaine public, la protection est refusée par les tribunaux "lorsque l'ordre de classement et la réunion des éléments n'ont donné lieu à aucun choix, à aucune manifestation de la personnalité parce que la nature de la publication imposait sans la moindre hésitation tels éléments et tels arguments à l'exclusion de tous autres".³ La nature factuelle des faits et le

1.- LAMY - Droit de l'informatique 1991, n°2303.

2.- Rapport du conseiller JONQUERES, R.D.P.I. 1986, n°84.

3.- DESBOIS Henri, *Le droit d'auteur en France*, Paris, 3e éd., 1978.

but recherché ne doivent pas commander le choix et le classement.

Le choix est une manifestation de liberté. Mais souvent cette liberté est celle de naviguer entre les contraintes ou simplement les nécessités et ces dernières sont plus ou moins pesantes selon la nature de l'œuvre.

La réunion d'éléments factuels (un ensemble de faits historiques) peut être la conséquence d'un choix fondé sur des critères objectifs. Pour qu'un choix soit considéré comme original, il faut que l'auteur ait disposé d'un minimum de liberté. L'œuvre pourra acquérir un caractère individuel alors même que les données sélectionnées appartiennent au domaine public.

Dans un arrêt datant du 20 mai 1980, la Cour de Cassation s'est récemment prononcée pour une protection d'un ouvrage sur la base du critère de sélection. "Constitue une contrefaçon le fait pour un auteur de reproduire dans son propre ouvrage un certain nombre de photographies figurant dans une autre œuvre, dès lors que ces photographies, bien que reproduisant des documents placés à la disposition du public à la Bibliothèque Nationale, constituaient une sélection de ces documents soigneusement opérée par l'auteur du premier ouvrage, sélection qui est un des éléments de l'originalité de l'ouvrage contrefait".¹

Outils relevant d'un savoir-faire, mais également créations de forme, les produits documentaires doivent pouvoir accéder à la protection, à condition d'être originaux.

1.- Cass; 1e civ., 20 mai 1980, JCP 1980, IV, 289.

Reconnaître une protection potentielle aux produits documentaires nous amène à admettre la protection d'une nomenclature de dictionnaire ou d'un lexique bilingue. Cette position est confirmée par un arrêt récent de la Cour d'Appel de Paris¹ qui a jugé qu'un dictionnaire informatique anglais-français constituait une œuvre de l'esprit et pouvait être susceptible de contrefaçon.

Le procès opposait la société Masson, éditeur d'un dictionnaire anglais-français et français-anglais d'informatique en deux volumes et la société Harrap's Limited.

La société Masson estimait que la majorité des nouvelles rubriques ou modifications apportées aux anciennes dans la dernière édition du dictionnaire Harrap's étaient reprises de la dernière édition de son dictionnaire bilingue ; cette copie était donc constitutive de contrefaçon.

La société Harrap's France faisait valoir, quant à elle, que le dictionnaire Masson ne pouvait bénéficier de la protection de la loi de 1957, en raison de la matière traitée ; le langage informatique, "de libre parcours et du domaine public" étant par nature insusceptible d'appropriation.

Pour établir la contrefaçon, malgré la présence obligée de certains termes techniques, la Cour d'Appel s'est basée sur le critère du choix ; elle a très précisément décrit les modalités de création d'un dictionnaire et a indiqué en quoi le choix pouvait être un facteur d'originalité : "La création d'un dictionnaire suppose tout d'abord la définition en qualité et en quantité, de la masse lexicale à retenir ; (...) l'auteur d'un

1.- Paris, 4e ch. Appel, 21 mars 1989, Lamy - droit de l'informatique, mise à jour F, page 6.

dictionnaire technique est amené à effectuer un premier tri parmi les termes courants (...) ; que cette première opération qui requiert culture et attention, orientée par la vision de l'œuvre à réaliser, doit être qualifiée d'opération de l'esprit ; qu'après le choix des mots à retenir ou à éliminer, intervient le parti qui sera pris ou de restreindre l'ensemble de l'œuvre dans un volume réduit., ou, à l'inverse, d'être aussi exhaustif que possible... ; que ces choix et les opérations intellectuelles qu'ils requièrent déterminent de nouvelles opérations intellectuelles qui, déjà, contribuent à donner à l'ouvrage, si technique soit-il, une originalité certaine, dès lors qu'il s'agit réellement, comme en l'espèce, d'un dictionnaire établi à partir de sources documentaires multiples et non d'une simple présentation, par ordre alphabétique, de listes de mots dont l'équivalence dans une langue étrangère aurait été définie par un tiers, telle par exemple la commission compétente pour l'informatique".

La Cour d'Appel décrit ici très précisément ce qu'est le programme d'information d'un dictionnaire et en quoi consistent les choix à opérer, ces choix représentant une opération intellectuelle. Les termes employés "opération intellectuelle", "de l'esprit" sont proches des notions de travail, effort intellectuel, retenues par les tribunaux en matière de compilation.

Il est à noter que l'argumentation de la Cour d'Appel peut s'appliquer à n'importe quel dictionnaire, qui par définition, est élaboré en fonction de choix définis par l'éditeur.

Quant à la traduction, la société Harrap's argumentait qu'un dictionnaire de langue ne constitue pas en soi une œuvre originale, au motif que les termes techniques se traduisent mot à mot.

A contrario, la Cour d'Appel retient que la société Masson opère des choix par suppression, remplacements ou ajouts de certains termes traduits et que les traductions elles-mêmes subissent des variations.

"Il n'a nullement été démontré que le dictionnaire réalisé par Ginguay (Masson) aurait été l'addition de traductions d'une banalité certaine ou appartenant au domaine public en raison de leur source ; que bien au contraire les variations qu'elles subissent (...) démontrent à elles seules, leur caractère d'originalité... Si certaines traductions, correspondant à un mot d'usage courant en informatique désignant un objet n'ayant qu'une seule dénomination dans chaque langue, apparaissent obligées (telle : trolley pour chariot à papier) et se retrouvent nécessairement dans chaque dictionnaire, certains mots dont la présence n'apparaît pas s'imposer à l'évidence dans un dictionnaire de l'informatique (*zip-code* : code postal) sont néanmoins ajoutés à l'édition Harrap's 1985 (...) que surtout d'innombrables traductions non obligées (*dead-lock* : étreinte fatale) et de nombreux ajouts aux rubriques apparaissent au Harrap's 1985 qui correspondent très exactement au texte de Masson 1981".

Les juges concluent à une copie "exercée de façon massive, (...) qui peut être relevée à chaque page du dictionnaire Harrap's, la matière reprise au dictionnaire Masson constituant entre le quart et le tiers du texte de la troisième édition du Harrap's".

La marque de choix subjectifs est également l'argument mis en avant par le président d'Utrecht, dans une décision du 17 août

1989¹, pour reconnaître une protection à un dictionnaire néerlandais ; lors de l'élaboration du dictionnaire, seuls certains mots de la langue avaient été retenus, en fonction de critères spécifiques.

L'examen de la doctrine et de la jurisprudence montre que, d'une façon générale, le critère du choix peut être retenu pour caractériser l'originalité d'une œuvre, à condition que le résultat de cette sélection soit suffisamment individuel.

2.3.2.- Les corpus de référence et les dictionnaires à usage d'automates

Les corpus de référence et les dictionnaires électroniques sont des collections structurées d'informations. Les corpus de références rassemblent des extraits de textes eux-mêmes protégés par un droit d'auteur. Les dictionnaires électroniques rassemblent des informations appartenant au domaine public.

Il s'agit de se demander alors si un corpus de références ou un dictionnaire électronique peuvent être protégés par un droit d'auteur et le problème de la protection de ces objets se pose indépendamment de leur contenu.

Cette question amène à rechercher sur quels éléments l'originalité peut porter.

Les tribunaux français reconnaissent depuis longtemps une originalité aux compilations d'éléments non protégés, comme les guides, les almanachs ou les annuaires. L'examen de quelques décisions significatives nous permettra de

1.- Lamy - Droit de l'informatique, 1991, n°2303.

comprendre pourquoi une activité de compilation peut être protégée par un droit d'auteur.

Un effort de recherche, mise en ordre, présentation des différents éléments confère à l'ouvrage considéré une singularité qui l'autorisera à jouir d'un droit privatif.

Toutefois, le fait d'accepter que les compilations soient susceptibles d'être protégées n'est rien d'autre qu'un point de départ car reconnaissance d'une simple potentialité. Il faut donc s'interroger sur le degré de protection et les limites que leur conférerait le droit d'auteur.

Il a été jugé que le catalogue général officiel d'une exposition (d'électricité), dressé d'après une classification méthodique et complète et à la suite d'un travail préliminaire sérieux, de combinaisons réfléchies, constituait une œuvre de l'esprit et de l'intelligence susceptible d'être protégée par les lois sur la propriété littéraire.¹

Inversement, un catalogue dont les nomenclatures furent établies par ordre alphabétique, méthode naturellement propre à la matière, ne put bénéficier de la protection légale.

Dans un arrêt du 4 mars 1982, la Cour d'Appel de Paris a relevé que des objets figurant dans un catalogue, qui fut ensuite reproduit, étaient groupés sur la page de droite, disposés sur un tissu, dans une composition recherchée de lignes et de volumes, que les spécifications étaient indiquées sur les pages de gauche, assorties entre autres de dessins humoristiques. L'ensemble de ces caractéristiques ne se retrouvant combinées dans aucun autre catalogue de produits

1.- Trib. civ. Seine 9 mai 1884, Ann. Prop. Ind., 1885, 155.

similaires (d'où la reconnaissance implicite de la nouveauté), la Cour a pu en déduire que l'ouvrage reproduit constituait une création de l'esprit, reflétant la personnalité de son auteur.¹

Comme pour les catalogues, des jugements anciens et des décisions plus récentes ont reconnu ou dénié à des guides le caractère d'œuvre de l'esprit.

Pour admettre l'originalité d'un "guide de la famille", les juges du Tribunal Correctionnel de la Seine ont employé les arguments suivants : "(...) l'expression générale "écrits en tout genre" s'applique à toute composition littéraire, quel que soit son degré de mérite ou d'utilité, (...) elle vise notamment le plan d'un ouvrage, son mode de division, alors même que cet ouvrage ne comporterait que des compilations ou des remaniements, du moment que les éléments en ont été disposés de telle manière qu'ils impriment à la publication un caractère nouveau (...) Vainement l'on objecterait que les éléments constitutifs d'un guide sont nécessairement empruntés au domaine public et que tout guide poursuivant le même but contiendra les mêmes indications (...) L'originalité de l'œuvre d'où découle sa protection consiste dans la classification des indications et des renseignements donnés ainsi que dans leur assemblage et leur présentation".²

Il a été récemment jugé qu'un guide pratique regroupant des renseignements administratifs faisant partie du domaine public constituait une œuvre protégée dès lors que la présentation, le choix et l'agencement de ces éléments portaient l'empreinte personnelle de leur auteur.³

1.- Paris 4e ch. B, 4 mars 1982, D. 83, I.R. 93

2.- Trib. Corr. de la Seine, 21 mars 1913, Droit d'Auteur, 1914, page 10.

3.- Paris, 4e ch. A, 6 mai 1987, D.S. 1988, Sommaire 202.

A l'inverse, ne peuvent bénéficier de la protection du droit d'auteur des fiches techniques qui ne sont que la reproduction à l'identique des données fournies par le client, selon un ordre imposé par celui-ci et dans une présentation banale et usuelle.¹

Les almanachs, répertoires, annuaires, calendriers ou indicateurs "qui constituent assurément le dernier degré de cette échelle qui commence en quelque sorte au ras de terre pour se perdre au plus haut des régions éthérées où plane le génie", peuvent également bénéficier de la protection de la loi, "si infimes, si misérables que ces productions de l'esprit nous paraissent".²

Elles accèdent à la protection par un choix, un classement et une présentation originale de la matière dont résulte généralement une facilité particulière pour la consultation de l'ouvrage. Celles-ci se distinguent des compilations qui comportent l'exposé d'un sujet ; est en général seulement repris un ensemble de renseignements matériels.

Mais une distinction entre les deux reste limitée car la protection est accordée aux unes et aux autres aux mêmes conditions et sur les mêmes arguments.

La Cour d'Appel de Paris a récemment décidé qu'un répertoire professionnel d'agence de conseils en publicité et de presse présentait un caractère de nouveauté et d'originalité dans la mesure où, à côté de renseignements figurant déjà dans l'annuaire des P et T, se trouvaient certaines informations relatives au mode d'intervention de l'agence, à sa forme, son

1.- Cass. civ. 1e, 8 décembre 1987, JCP 1988, IV, 67.

2.- POUILLET Eugène, *Traité théorique et pratique de la propriété littéraire et artistique*, Editions Marcal et Billard, Paris, 1908, page 51.

chiffre d'affaires, ses responsables...¹ Les arguments invoqués dans cette décision n'intéressent même plus la forme de l'ouvrage mais son contenu ; on assiste à un glissement de la forme vers le fond et en conséquence une protection du contenu informationnel.

Pour terminer ce panorama jurisprudentiel, on peut citer trois décisions contradictoires, qui concernent les annuaires.

La première décision a, par un attendu général, dénié toute protection au motif que "consistant dans un assemblage de matériaux fournis par le domaine public, ils (les annuaires) ne sont point à ce titre susceptibles d'une propriété privative assimilable à la propriété littéraire".²

Dans la deuxième décision, le refus de protection d'un annuaire de l'alimentation a été argumenté ainsi : "il s'agit d'une forme d'arrangement et de composition, s'imposant par la force même des choses, il s'agit en effet de former des rubriques concernant le genre ou les espèces des diverses données alimentaires, de les placer par ordre alphabétique, de manière à former les principales divisions de l'ouvrage... ; que ces classifications ont entre elles, pour un annuaire de l'alimentation ou de l'industrie alimentaire, une ressemblance nécessaire dépendant de la nature même des choses".³

La troisième décision accorde une protection à l'annuaire de la société Didot-Bottin, mais la généralité des termes employés par les juges du fond restreint la portée de l'arrêt. "Sans doute, les noms, adresses et les divers renseignements dont se compose un annuaire tel que celui de la société Didot-Bottin

1.- Paris 4e ch., 19 juin 1984, D.S. 1985, I.R. 309.

2.- Paris 2 avril 1896, Ann. Prop. Ind., 1898, 62.

3.- Trib. civ. Seine, 31 mars 1908, Paris.

sont de notoriété courante et appartiennent comme le dit le Tribunal, au domaine public ; mais le fait de les réunir en un ensemble ordonné et complet et de les classer en diverses nomenclatures comporte un travail important ; le résultat de ce travail, comme celui de toute industrie humaine, donne lieu à un droit privatif".¹

A la lecture de ces décisions, plusieurs remarques s'imposent :

- la marge de manœuvre des juges pour accorder ou refuser la protection se révèle très étroite. Mais il ne semble guère possible de démontrer l'originalité d'une compilation autrement qu'en invoquant le caractère ingénieux de la classification et de la présentation ou de réfuter l'originalité en arguant de la présentation banale, usuelle ou propre à la matière.

- les références à l'empreinte de la personnalité de l'auteur ont pratiquement disparu et sont remplacées par des notions telle que le travail (personnel, de classement, important) ou l'effort (intellectuel, de recherche). C'est avant tout une performance intellectuelle qui est prise en considération plutôt que la reconnaissance d'un résultat tangible à travers une œuvre individualisée.

- Le critère de nouveauté est parfois invoqué en complément de celui d'originalité. Bien qu'il existe, comme nous l'avons vu, une différence de principe, une confusion est souvent faite entre le concept d'originalité et celui de nouveauté. Ces deux notions sont si proches que l'on glisse insensiblement de l'une à l'autre. "A l'instar du juge des brevets, le juge est tenté de

1.- Paris 18 décembre 1925, Ann. Prop. Ind. 1926, 99.

situer l'œuvre de l'esprit dans l'ensemble de la production d'une époque ou d'une école en superposant nouveauté et originalité, deux concepts qui à bien des égards se recouvrent".¹

a) L'originalité des corpus de références

La description détaillée d'un corpus de références dans la première partie de cette thèse fait clairement apparaître une subjectivité non seulement dans le choix, mais aussi dans la structuration des données. L'apport intellectuel est indéniable. Un corpus de références pourra être protégé par le droit d'auteur, au même titre qu'une base de données bibliographiques ou lexicographiques.

Mais admettre la protection éventuelle d'un corpus de références ne doit pas faire oublier que la matière même du corpus (les extraits d'ouvrages, de journaux...) est éventuellement protégée par un droit d'auteur. L'autorisation de reproduction d'œuvres pré-existantes ou d'extraits d'œuvres doit faire l'objet d'une demande préalable auprès de l'auteur ou de ses ayant-droits.

b) L'originalité des dictionnaires électroniques

Les dictionnaires et grammaires électroniques ont pour objet de recenser tous les mots de la langue et de les organiser en dictionnaire. Ces systèmes d'information sont entièrement nouveaux par rapport aux ouvrages traditionnels ; ils s'en distinguent par leur contenu et leur structure.²

1.- Rapport du conseiller JONQUERES, R.D.P.I., 1986, n°54 et n°79.

2.- Voir la première partie de cette thèse.

Décrire la langue revient à en construire une description formelle dans laquelle toute information de nature sémantique doit être écartée car non compréhensible par la machine.

A la différence des dictionnaires usuels, destinés à être utilisés par des humains, les dictionnaires électroniques contiendront des informations morpho-syntaxiques, définies comme objectives et appartenant au domaine public. Ces données doivent être totalement explicites et aussi complètes que possible.

Cette double exigence de description objective et d'exhaustivité¹, requise pour les dictionnaires à usage d'automate, pose avec une acuité accrue le problème de l'application du critère d'originalité. Nous pouvons formuler ce problème en reprenant la formulation de Maître Jean Martin : "quels éléments constitutifs du système d'information convient-il de prendre en considération pour déterminer s'il y a œuvre ?"²

Un outil de description de la langue, comme un texte législatif est à la fois figé, car expression d'une norme, et évolutif car résultat d'une évolution dans un consensus; un mécanisme descriptif de la réalité en même temps que d'occurrence de cette réalité. C'est pourquoi la description d'une langue naturelle ne pourra jamais être totalement objective (à cause de l'infinitude potentielle des significations et des contextes d'occurrences), ni totalement exhaustive (à cause de l'infinitude potentielle des formes du langage naturel).³

1.- Un dictionnaire électronique n'est pas réalisé en fonction d'un programme d'information.

2.- MARTIN Jean, *Quelques aspects des enjeux* - In : banques de données et droit d'auteur - Colloque IRPI, Paris, 27 novembre 1986, LITEC, 1987, page 7.

3.- DEGREMONT Jean-François, *Ethnométhodologie et innovation technologique : le cas du traitement automatique des langues naturelles* - Doctorat d'Ethnologie - Université Paris VII, 1989.

Mais cette absence d'objectivité et d'exhaustivité tient uniquement à la nature du langage naturel ; par définition, dans la réalisation d'un dictionnaire électronique, les données ne sont pas volontairement sélectionnées. "Survient alors ce qu'on peut appeler le "paradoxe de l'exhaustivité : si on a tout intégré dans la banque de données, peut-on encore parler de subjectivité ?".¹

Une sélection peut être synonyme d'originalité. En revanche, doit-on écarter de la protection une banque de données qui prétend être exhaustive dans un domaine ?

Un arrêt ancien a reconnu une protection à un annuaire du fait que l'entreprise avait réalisé un ensemble ordonné et complet des adresses, les avait classées en les répartissant en plusieurs nomenclatures.²

Plus récemment les juges du fond ont admis que revêtait un caractère d'originalité un ouvrage, qui loin d'être un simple assemblage non structuré de connaissances éparses, présente l'intérêt de faire une étude, aussi complète que possible, des carences en oligo-éléments chez les ruminants : en raison de son caractère synthétique (travail sur la matière), cet ouvrage est marqué par la personnalité de celui qui a réalisé cette synthèse des connaissances scientifiques en la matière, le caractère d'œuvre de l'esprit ne pouvant lui être dénié³. Le compilateur peut tout dire, sans faire œuvre créatrice ; il ne deviendra auteur qu'à la condition que l'information soit synthétisée et structurée.

1.- GINSBURG Jane, *Banque de données et droit d'auteur aux Etats-Unis* - In : *Banques de données et droit d'auteur - Colloque IRPI, Paris, 27 novembre 1986*, LITEC, 1987, page 105.

2.- Paris, 18 décembre 1925, Ann. prop. ind. 1926, 99.

3.- Paris, 20 avril 1989, D. 1989, I.R. 177.

Un dictionnaire électronique possède une structure qui lui est propre et peut varier selon les écoles scientifiques qui proposent des modèles de description morpho-syntaxiques différents. Mais la structuration d'un dictionnaire à usage d'automate va résulter non seulement de l'organisation des données (un dictionnaire impliquant obligatoirement une notion d'ordre, donc de fichier), mais aussi de l'utilisation du logiciel qui permet d'effectuer des traitements divers et d'offrir des combinaisons multiples.

Les dictionnaires électroniques, comme les bases de données bibliographiques ou lexicographiques, échappent "par définition à la rigidité" et ont une "disposition à géométrie variable"¹. Ainsi contrairement au dictionnaire usuel où la structure de la fiche lexicographique est normalisée et immuable tout au long de l'ouvrage, le dictionnaire électronique peut présenter à l'utilisateur des formes extrêmement variées, tant par l'ordre dans lequel sont présentés les mots et leurs caractéristiques, que par la forme des présentations.

Il faudrait, dans le cas des systèmes électroniques d'informations, prendre en compte "les éléments structurants, voire relationnels, des données..., qui sans ordre apparent sont en réalité dans une potentialité d'ordres multiples précis et voulus comme résultats d'une activité créatrice"².

Un corpus de références ou un dictionnaire électronique, qui ne sont pas que de simples archivages d'information, peuvent ainsi constituer des œuvres protégeables.

1.- VIVANT Michel, *Informatique et propriété industrielle*, J.C.P. 1984, Ed. G.-I-3169, n°22.

2.- MARTIN Jean, *Quelques aspects des enjeux* - In : Banques de données et droit d'auteur - Colloque IRPI, Paris, 27 novembre 1986, LITEC, 1987, page 7.

Troisième partie

Le statut juridique des produits issus des industries de la langue

3.- Le statut juridique des produits issus des industries de la langue

"La soumission du logiciel à des dispositions particulières issues (...) de la loi du 3 juillet 1985 (copie non autorisée, durée de protection plus courte, droits dévolus à l'employeur...) oblige à un effort de délimitation. Seul ce qui est proprement logiciel relèvera de la loi de 1957-1985, et tout ce qui ne sera pas distinctement, proprement, logiciel restera soumis à la loi de 1957 dans sa forme de droit commun".¹

Ces problèmes de frontières entre deux statuts juridiques concernent à la fois l'environnement immédiat du logiciel (la documentation associée au logiciel et le "*look and feel*") et les produits informatiques complexes composés d'au moins deux éléments qui peuvent être distincts ; l'un semblant relever du droit des logiciels (loi de 1985), l'autre semblant relever du droit commun.

"Il n'est donc pas incongru d'attendre *a priori* un régime juridique éclaté (le logiciel... et le reste) quand la raison, au moins économique, voudrait que ces créations soient chacune appréhendées comme un tout. Economiquement c'est bien (...) un correcteur orthographique ou un système de traduction assisté par ordinateur, (...) qui va être commercialisé et non point un conglomerat d'éléments pour partie logiciels, pour partie d'une autre nature"². Parmi les produits fabriqués dans le cadre des industries de la langue, nombreux sont ceux à propos desquels se posent ces questions de statut juridique³.

1.- Lamy droit de l'informatique 1991, n°1305.

2.- Lamy droit de l'informatique 1991, n°1701.

3.- POITEVIN Christine, *Un cadre juridique harmonisé pour les langues européennes*, In : Les industries de la langue, enjeux pour l'Europe - Actes du colloque de Tours, Revue Encrages, Paris, n° 16, 1986, page 78.

**Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques**

Relèvent-ils uniquement de la loi de 1957, uniquement de la loi de 1985 ou simultanément des deux régimes, ainsi que le proposent certains juristes¹.

Dans ce cas, "il faudra, simplement, opérer une ventilation entre ce qui relève de la loi de 1985, dans la mesure où elle déroge au droit commun de la propriété littéraire et artistique, et ce qui demeure protégé par la loi de 1957".

Ce raisonnement conduit à admettre un régime juridique éclaté pour des objets tels qu'un correcteur orthographique, en attribuant par exemple la propriété de la partie logicielle à l'entreprise et la propriété de la partie non logicielle (les dictionnaires) éventuellement au salarié de l'entreprise ou à un créateur indépendant.

La réponse au problème du statut juridique ne peut être fournie qu'après s'être demandé si les produits issus des industries de la langue, considérés dans leur globalité et non à travers chacune de leurs composantes, peuvent être qualifiés à des logiciels. En cas de réponse positive, la loi du 3 juillet 1985 leur sera appliquée. Dans le cas contraire, le statut légal de ces produits relèvera de la loi du 11 mars 1957.

Le choix entre deux statuts juridiques pour les produits issus des industries de la langue ne pourra donc être fait qu'après avoir répondu à la question : qu'est-ce qu'un logiciel ?.

Remarquons que "tant que la question se posait par référence à la loi de 1957, il ne s'agissait pas de savoir si on était en présence ou non d'un logiciel, mais d'une œuvre au sens de la

1.- HUET Jérôme, *La protection des logiciels, pratiques passées, régime actuel et difficultés futures*, in *Expertises*, n° 86, Août 1986, page 174.

loi et satisfaisant à ses exigences propres. A la limite, peu importait qu'il s'agisse ou non d'un logiciel"¹. A partir du moment où il existe des règles dérogatoires au droit commun de la propriété littéraire et artistique, une définition du logiciel devient "indispensable pour la formulation de la règle de droit"².

Dans ce chapitre, nous allons nous efforcer de préciser les critères qui permettent de désigner des objets par le mot "logiciel". Nous appliquerons ensuite ces critères aux objets étudiés dans cette thèse afin de déterminer leur statut légal.

3.1.- Contribution au débat sur la définition du logiciel

3.1.1.- L'élaboration des définitions du mot "logiciel"

3.1.1.1.- Les définitions dans les textes officiels

Le texte officiel le plus récent, qui est la Directive du Conseil des Communautés Européennes concernant la protection juridique des programmes d'ordinateur, adoptée le 14 mai 1991, ne fournit pas de définition du terme "logiciel".

En droit français, bien que la loi du 3 juillet 1985 emploie le terme "logiciel" à plusieurs reprises, en particulier dans son titre V, elle ne donne aucune définition implicite ou explicite du mot. Il n'existe donc pas, dans notre pays, de définition

1.- VIVANT Michel, Intervention au colloque "Banques de données et droit d'auteur", Colloque IRPI, Paris 27 novembre 1986, Librairies Techniques, 1987, page 41.

2.- KEREVER André, *Protection par le droit d'auteur ou protection "sui générés" ?* - In : Banques de données et droit d'auteur - colloque IRPI, Paris, 27 novembre 1986, Librairies Techniques, page 78.

légale du terme "logiciel". On doit cependant, pour être tout à fait complet, préciser qu'un texte officiel fournit une définition technique de ce qu'est un logiciel. Ce texte, à vocation linguistique, a été rédigé et publié dans le cadre des travaux des Commissions nationales de terminologie¹.

Ces commissions ont pour objet principal de soutenir l'emploi de la langue française, y compris dans les domaines techniques où l'anglais risquait de devenir la langue de travail obligée par le fait de l'inexistence de termes français utilisables. Les Commissions nationales de terminologie, qui réunissaient des spécialistes de la langue et des spécialistes des domaines considérés, ont publié par voie d'arrêté des listes de termes et de définitions. Le volume de ces listes est assez faible (quelques dizaines de termes par an) en regard du volume de la terminologie technique (environ 4 millions de termes techniques sont utilisés quotidiennement dans les différents domaines scientifiques et industriels). L'emploi des termes publiés est théoriquement une obligation légale dans certains cadres en vertu de plusieurs textes : on devrait ainsi ne plus trouver que les termes autorisés dans les notices techniques des équipements vendus en France, ainsi que dans tous les textes des contrats passés avec l'Etat. Bien entendu le législateur est le premier à constater et à affirmer que "la langue ne se décrète pas" et que l'obligation légale d'emploi vise en réalité à forcer la mise en place d'habitudes qui feraient passer le terme nouveau dans le langage courant.

Cet effort a d'ailleurs été couronné de succès dans de multiples cas (le mot "bureautique" en est un exemple célèbre), mais de très nombreuses propositions sont soit des demi-

1.- Ces commissions ont été créées par décret n° 72-19 du 7 janvier 1972, relatif à l'enrichissement de la langue française (J.O. du 9 janvier 1972).

succès (le terme "mercatique" en lieu et place de l'anglicisme "marketing"), soit des échecs (le terme "baladeur" en lieu et place de "walkman").

L'arrêté du 22 décembre 1981, relatif à "l'enrichissement du vocabulaire de l'informatique"¹, définit le logiciel comme étant : "l'ensemble des programmes, procédés et règles, et éventuellement de la documentation, relatifs au fonctionnement d'un ensemble de traitement de données".

Ce texte a été vivement critiqué par certains juristes. Ainsi Jean-Michel Piot², affirme que cette définition est imprécise et qu'elle "est loin d'enrichir la langue française". Il ajoute "qu'une telle imprécision est dommageable quant à l'interprétation de la loi de 1985 qui se réfère à ce terme sans le définir". L'exploitation de ces définitions par des spécialistes du droit est effectivement difficile mais ce n'est pas pour les raisons invoquées par l'auteur, à savoir "l'imprécision" des travaux des Commissions de terminologie et éventuellement "le manque de compétences" lexicographiques ou techniques des membres de ces commissions.

Plusieurs raisons permettent d'expliquer pourquoi le résultat du travail des Commissions nationales de terminologie est difficile à utiliser par les juristes. Ces raisons découlent directement de la mission des Commissions, du mode de sélection des termes sur lesquels elles travaillent et des procédures mises en œuvre pour produire des définitions pour les termes nouveaux proposés.

1 JO(NC) du 17 janvier 1982, page 624.

2.- PIOT Jean-Michel, *Les créations informatiques complexes* - Mémoire de D.E.A. , Montpellier, 1987.

La première de ces raisons est que l'objet même du travail des Commissions se place dans le cadre de la politique linguistique de l'Etat français (promouvoir et encourager l'emploi de la langue française en respectant les diversités culturelles). Ce respect des diversités culturelles a pour conséquence de reconnaître l'existence de variantes potentiellement infinies : le français du Québec, le français d'Haïti, etc.

Ce respect des diversités va jusqu'à la personne physique : la définition de "langue française" est toujours restée générale et n'a jamais pris la forme d'une contrainte absolue et définie avec une entière précision. On serait d'ailleurs bien incapable pratiquement de mettre en place cette contrainte puisqu'elle supposerait l'existence d'un dictionnaire de référence dont il est démontré depuis le début des années 1950 qu'il n'est pas possible de le réaliser pour une langue naturelle¹. On trouvera ainsi des textes en droit français faisant obligation, dans certains cas, d'utiliser ou de connaître la "langue française", mais cette obligation ne sera jamais accompagnée d'une description exhaustive de la "langue française" qui formerait la référence permettant d'appliquer des mesures coercitives².

La seconde raison découle de la première. En cherchant à proposer des termes qui vont "naturellement" s'imposer dans la société, les Commissions nationales de terminologie s'interdisent de fournir des termes et des définitions dont la complexité ou la technicité provoqueraient un rejet. Il y a, dans le travail des Commissions, la recherche d'un équilibre

1.- voir à ce sujet, dans la première partie de cette thèse, le chapitre qui traite de la langue naturelle, de son infinitude potentielle des sens, ainsi que de son caractère permanent d'évolution.

2.- Ces textes sont regroupés, entre autres, dans DE BESSE Bruno, *Termes techniques nouveaux*, Feutry Editeur, Paris, 1982.

perpétuel entre la simplicité et la précision, dans une position qui s'apparente fortement à la tentative des encyclopédistes du dix-huitième siècle lorsqu'ils cherchaient à être compris du plus grand nombre.

Bien que les membres des Commissions, en "honnêtes hommes", s'efforcent de rester le plus précis possible, d'un point de vue technique, l'unanimité de l'acceptation par les usagers prime sur la clarté du sens proposé pour les néologismes. La recherche de cet équilibre peut être en contradiction, comme le devenir de nombreux termes l'a clairement montré, avec certains usages du résultat des travaux des Commissions. Ainsi, pour ce qui est du terme "logiciel", la définition proposée par la Commission nationale de terminologie de l'informatique dans l'Arrêté du 22 décembre 1981, présente l'avantage d'être facilement compréhensible. Cette définition, par contre, n'est pratiquement pas utilisable par un juriste qui chercherait à savoir si un objet qui lui est présenté est ou n'est pas un logiciel et entre ou n'entre pas dans le domaine d'application de la loi de 1985.

En proposant un terme et sa définition, les membres de Commissions ne cherchent rien d'autre qu'à s'efforcer de clarifier une situation linguistique potentiellement ou factuellement confuse.

La troisième raison qui rend le travail des Commissions peu utilisable par les juristes est leur mode de choix des termes qui vont être définis. Les Commissions de terminologie ne travaillent que sur la base de constats linguistiques puisés principalement dans la littérature de vulgarisation scientifique et technique et dans la presse quotidienne. Ainsi, pour un terme dont l'emploi est déjà constaté dans la littérature technique ou de vulgarisation, les membres des

Commissions vont fournir une définition dont il leur semble qu'elle recouvre le maximum d'emplois et qu'elle évitera les faux-sens aux rédacteurs professionnels (en particulier les journalistes et tous les intermédiaires entre les milieux spécialisés et le grand public). Pour un terme d'origine anglo-américaine dont l'usage potentiellement massif est constaté, ils proposeront un équivalent dont les consonances et la morphologie seront plus "conformes" à la tradition latine ou grecque de la "langue française". La définition qui en sera donnée aura également pour objet de fournir une sorte de norme aux usagers francophones.

A aucun moment, les Commissions n'ont pour mission ni pour objectif de fournir un inventaire exhaustif de la langue française. Elles font un choix très sélectif parmi les très nombreuses occurrences de termes nouveaux, que ces termes soient empruntés à d'autres langues ou qu'ils aient une apparence francophone. En particulier, leur critère de choix n'est pas de prendre la totalité des termes utilisés dans les textes du droit français.

Il n'y a donc pas lieu de chercher une cohérence juridique dans les résultats des travaux des Commissions nationales de terminologie, pas plus qu'il n'y a lieu de chercher une cohérence systématique entre les travaux des différentes Commissions de terminologie. On se souviendra, à titre d'exemple, des propositions publiées au JO par deux Commissions nationales de terminologie pour remplacer le terme "gas-oil". L'une avait publié le terme "fioul" et l'autre le terme "fuel".

La quatrième raison qui rend le travail des Commissions nationales de terminologie peu utilisable par des juristes découle du processus mis en œuvre dans l'élaboration des

définitions. Constatant des usages divergents d'un même terme ou l'usage d'un terme emprunté à une langue étrangère, les membres des Commissions s'efforcent de proposer le plus rapidement possible une définition stabilisatrice ou un équivalent français. Ils travaillent dans des domaines dont la caractéristique principale est souvent d'évoluer très rapidement au plan technique (les Commissions sont réparties par thèmes : aéronautique, informatique, télédétection...) et ils ne cherchent pas à faire œuvre définitive ni même œuvre à long terme dans les définitions qu'ils proposent car ils savent que leur domaine de travail aura changé fortement dans un délai bref. Ils cherchent à stabiliser, à préciser dans l'instant, le terme sur lequel ils travaillent. Et leur recherche est celle du consensus de l'instant, cet instant se mesurant en mois plus qu'en années.

En effet, si leur action aboutit, c'est à dire si le terme et sa définition sont acceptés par les usagers, c'est bien parce qu'ils ont su trouver le consensus avant que ne se soit imposé le terme étranger. La lecture des listes de propositions des commissions est, à ce titre instructif. Les propositions qui portent sur des équivalents à des termes étrangers qui étaient déjà fortement implantés ne se sont souvent pas pratiquement imposées, alors que les propositions qui sont venues très rapidement après l'apparition du concept dans la littérature ont souvent été adoptées par le corps social. Cet impératif de rapidité est parfois contradictoire, dans les domaines en évolution rapide, avec celui de pérennité comme avec celui de précision des définitions. Une recherche de plus grande précision engendrera des débats longs et inutiles puisque, lorsque le terme sera publié, il ne s'imposera pas (et aura probablement évolué en échappant à la précision "officielle"). Une recherche de pérennité engendrera, outre des débats également longs, une rédaction prudente, globalisante,

générale de la définition qui débouchera sur un manque de précision, contradictoire avec l'objectif stabilisateur de la Commission.

Il convient enfin de faire une remarque sur la composition des Commissions nationales de terminologie. Les Commissions nationales de terminologie sont constituées de membres de droit et de membres experts. Les membres de droit sont issus des ministères et organismes publics qui ont tutelle ou compétence administrative dans le domaine considéré. Les textes qui définissent la composition des Commissions ne précisent pas le mode de désignation de ces membres de droit dans leur organisme d'origine, ce qui amène une grande irrégularité dans les compétences des personnes physiques qui sont effectivement désignées. Les autres membres des Commissions sont volontaires et non rémunérés. Ce sont, le plus souvent, des spécialistes du domaine considéré qui n'ont généralement pas de compétences linguistiques ni juridiques, bien que certains d'entre eux aient l'expérience de la rédaction lexicographique (enseignants, personnels de l'AFNOR, etc.). La compétence linguistique et la coordination entre les travaux des différentes Commissions sont théoriquement apportées par les membres de droit issus de structures de coordination comme le Commissariat Général de la langue française, rattaché aux services du Premier Ministre.

Les Commissions ont cependant un rythme de travail et de publication très fortement dépendant de la personnalité des Présidents. Leur productivité a varié de zéro à une centaine de termes par an et par Commission entre 1979 et 1989. En accord avec leur objet, ces structures sont très souples et mouvantes. Il ne faut donc pas attendre d'elles une cohérence linguistique analogue à celle que l'on peut trouver dans les grands ouvrages lexicographiques élaborés par des

professionnels rémunérés et publiés dans un environnement où la concurrence commerciale élimine impitoyablement les produits défaillants.

Au vu des raisons évoquées ci dessus on aura compris que l'utilisation des travaux des Commissions nationales de terminologies à d'autres fins que celles qui correspondent à leurs missions, ne peut qu'être marginale et faite aux risques et périls de l'utilisateur. C'est en particulier le cas pour qui tenterait une exploitation juridique des définitions proposées par ces Commissions. Comme on pourra le constater tout au long de cette partie, une exploitation de cette définition est néanmoins possible à condition de conserver à l'esprit le contexte dans lequel cette définition a été rédigée et de ne pas hésiter à remonter aux principes théoriques qui ont guidé ses rédacteurs. Le plus important de ces principes, pour ce qui concerne la définition du mot "logiciel", est la théorie des grammaires formelles.

Puisqu'il n'y avait pas de raison particulière à se limiter à la seule définition à caractère officiel du mot "logiciel", nous avons mené une recherche plus large, en particulier dans les dictionnaires encyclopédiques destinés au grand public.

Notons, avant de discuter ici quelques unes de ces définitions, que les définitions non françaises du mot "logiciel" sont très proches des définitions françaises. Ainsi la loi américaine¹ définit le logiciel comme étant "un programme d'ordinateur et une suite d'instructions destinées à être utilisées directement ou indirectement dans un ordinateur pour obtenir des résultats".

1.- Loi de 1976 sur le droit d'auteur telle qu'amendée en 1980. Pour d'autres définitions voir le Lamy droit de l'informatique 1992, n° 1306.

3.1.1.2.- Les définitions dans les dictionnaires

Quatre définitions du mot "logiciel" ont retenu plus particulièrement notre attention :

a) Dictionnaire alphabétique et analogique de la langue française de 1977 (Le "Littré") : "ensemble des travaux de logique, d'analyse, de programmation, nécessaires au fonctionnement d'un ensemble de traitement de l'information (opposé à Matériel).

b) Grand Larousse en 5 volumes (1985) : "ensemble des programmes, procédés et règles, et éventuellement de la documentation, relatifs au fonctionnement d'un ensemble de traitement de données (par opposition au Matériel).

c) Grand Robert en 9 volumes (1985) : "ensemble des programmes, procédés et règles, et éventuellement de la documentation, relatifs au fonctionnement d'un ensemble de traitement de l'information (opposé à Matériel).

d) Dictionnaire des Sciences édité par Hachette (1990) : "programme(s), procédures, règles et éventuellement documentation nécessaires à l'exécution d'un traitement d'information par un ordinateur".

Il est intéressant de constater la très grande continuité des définitions depuis celles du Littré qui a quatorze ans, jusqu'à celle du dictionnaire des Sciences publiée en 1990. Cette continuité se retrouve également dans plusieurs autres termes tels que "organigramme" ou "algorithme", dont on trouvera les définitions dans les documents annexes de cette thèse.

Tout comme la définition de la Commission nationale de terminologie de l'informatique, ces définitions, qui en sont très proches, ne permettent pas de progresser d'un point de vue juridique. Bien au contraire, ainsi qu'on peut le remarquer, la définition du Grand Larousse 1985, du Grand Robert 1985, et du dictionnaire des Sciences 1990, sont purement et simplement reprises de la définition à vocation linguistique de la Commission nationale de terminologie de l'informatique.

La seule différence dans la phrase est son dernier mot "ensemble de traitement de données", qui est parfois remplacé par "traitement de l'information" ou par "ordinateur". Cette différence est non significative et on peut considérer qu'elle résulte d'une tentative, très habituelle dans les dictionnaires scientifiques et techniques comme dans les dictionnaires non spécialisés, de limitation de reprise trop servile du texte qui est évidemment à l'origine de la définition.

Il est d'ailleurs intéressant de constater que la présence de ces micro-transformations, qui visent à changer la forme mais non le sens de la définition, est directement liée à la technicité des mots définis ainsi qu'aux compétences techniques de l'équipe de rédaction des dictionnaires. La copie servile devient fréquente lorsque les compétences techniques des lexicographes sont voisines de zéro.

Ainsi un organigramme est un objet dont la définition peut prendre des formes très différentes. Il convient, en particulier, d'être attentif à bien distinguer, d'une part, l'algorithme qui est l'idée, la méthode de résolution du problème avec une connotation de généralité et, d'autre part, l'organigramme qui est une représentation graphique de l'algorithme. Il existe bien d'autres façons de représenter un algorithme avec une connotation de contextualité : un

organigramme peut servir à représenter un algorithme général, mais il est le plus souvent utilisé dans le cadre de l'application d'un algorithme général pour un cas particulier et comme une activité préparatoire à la programmation proprement dite.

Bien entendu, ces subtilités ont échappé à de nombreuses équipes de rédaction de dictionnaires qui, outre qu'elles se sont copiées les unes les autres, n'ont même pas, faute de savoir-faire technique, évité de se recopier servilement. Ainsi la définition du Grand Robert en 9 volumes de 1985 ("représentation graphique des sous-ensembles d'un système et des relations qui les lient entre eux") est la stricte reprise de la définition du Littré de 1977.

A cette occasion, on ne peut s'empêcher de fortement relativiser le rôle novateur que s'attribuent volontiers, au plan linguistique, les éditeurs de dictionnaires. Si l'on comprend parfaitement l'intérêt commercial que les éditeurs de dictionnaires peuvent avoir à présenter leurs ouvrages comme novateurs au plan linguistique, il est également bien clair que pour des raisons financières, les éditeurs n'auront que rarement les moyens de faire effectivement ce qu'ils annoncent. Quand on connaît le nombre d'heures et la multiplicité des compétences que demande la rédaction d'un mot qui recouvre un concept un tant soit peu une technique, on comprend qu'il n'est pas question ici de mettre en parallèle une activité lexicographique, qui relève nécessairement du volontariat, et l'activité éditoriale d'entreprises nécessairement contrainte par des considérations économiques.

En dehors même de ces considérations économiques, une autre raison pratique explique le fait que les éditeurs de

dictionnaires se comportent, dans les domaines techniques, plutôt comme des chambres d'enregistrement que comme des structures d'innovation lexicographique. Cette raison est, qu'à l'exception de très rares encyclopédies, comme l'Universalis, les définitions ne sont pas rédigées par des spécialistes du domaine mais par des lexicographes professionnels constitués en équipes de rédaction par l'éditeur et salariés avec des niveaux de rémunération notoirement peu élevés. Ces rédacteurs possèdent, sauf très rare exception, une formation littéraire, ce qui les amène à privilégier la forme rédactionnelle des définitions au détriment de leur fond. Bien entendu, de nombreux dictionnaires et encyclopédies destinés au grand public réussissent le plus souvent à marier la qualité de la rédaction avec la pertinence des définitions.

On pourrait penser que, ne pouvant se satisfaire de la concision des définitions trouvées dans la majorité des dictionnaires destinés au grand public, les juristes seraient contraints à se nourrir des définitions, souvent jargonantes, qu'ils extrairaient d'ouvrages rédigés par des spécialistes pour des spécialistes. Ceci pourrait excuser certains monstrueux contresens techniques commis dans des textes rédigés par des juristes. Notre expérience documentaire, en particulier pour ce qui concerne les concepts rattachés à l'informatique, infirme cette hypothèse. Il ne nous est en effet jamais arrivé de ne pas trouver, au moins dans un dictionnaire encyclopédique destiné au grand public, une définition présentant toutes les qualités requises pour un travail juridique.

Nous détaillerons plus loin, la différence fondamentale qui existe entre un langage formel et un langage naturel. La compréhension de cette différence est essentielle pour qualifier juridiquement les objets informatiques complexes tels que les systèmes experts, les correcteurs

orthographiques, les didacticiels. Cette différence est remarquablement expliquée dans l'article de Jacques Hébenstreit sous l'entrée "informatique" de l'Encyclopédie Universalis. De même, une définition tout à fait précise, en même temps que concise, de ce qu'est un algorithme est donnée dans le Grand Robert.

L'absence de compréhension par certains juristes de concepts importants des domaines dont ils ont à traiter, découle moins, nous semble-t-il, de la complexité du domaine que de l'incomplétude de la procédure documentaire que ces juristes auraient dû réaliser préalablement à leur travail. Cette considération serait sans importance si ses conséquences restaient confinées dans les armoires de la littérature grise.

Parfois cependant des publications juridiques réputées pour leur sérieux reprennent par inadvertance des textes qui reposent sur ces confusions techniques majeures. Nous en verrons un exemple plus loin lorsque nous traiterons du statut des systèmes experts¹.

3.1.2.- Le choix des mots qui composent les définitions

Jean-Michel Piot² se demande pourquoi dans la définition de la Commission de terminologie de l'informatique, est utilisé le

1.- Si la base de connaissances d'un système expert était écrite en langage naturel, comme l'affirment Jean Gasnault et Jean-Christophe Vignal, dans leur article intitulé "Réflexions sur le statut des systèmes-experts", alors les ordinateurs seraient capables de performances langagières comparables à celles d'un humain, ce qui n'est pas le cas en 1992. (article publié dans le Lamy droit de l'informatique 1987, mise à jour, page 3).

2.- PIOT Jean-Michel, *Les créations informatiques complexes* - Mémoire de D.E.A. , Montpellier, 1987.

terme "ensemble de traitement de données" alors que la quasi totalité des autres définitions utilise le terme "ordinateur" ou plus exactement le terme anglais "*computer*" qui pourrait aussi se traduire par calculateur.

Pour cet auteur, l'utilisation du vocable "ensemble de traitement de données" introduit une confusion par le fait qu'il est très large : "il apparaît totalement inapproprié car trop vaste. Un télex ou même un magnétophone traite des données". Il suggère de compléter le vocable "traitement de données" par le qualificatif "informatique".

Notre avis est tout à fait différent. A notre sens, ce n'est pas la définition française qui est trop large, mais ce sont les définitions étrangères qui sont très restrictives. Nous allons nous en expliquer plus loin mais, au préalable, focalisons notre attention sur la structure commune, d'un point de vue lexical, à toutes les définitions du terme logiciel.

3.1.2.1.- Une définition impressionniste

Les définitions du terme "logiciel" que l'on a pu trouver dans la littérature ont la forme d'une juxtaposition de termes dont chacun, pris séparément, désigne des objets et des concepts plus larges que les simples logiciels. Ces termes sont, dans la définition de la Commission nationale de terminologie de l'informatique, les suivants :

- *programme* : le terme de programme donne une idée très générale de séquences ordonnées d'opérations à réaliser, sans que ces séquences soient nécessairement des œuvres de langage ;

- *procédé* : ce terme donne une information légèrement différente en ce sens qu'il est habituellement utilisé pour des séquences d'opérations ayant pour but de produire des objets (de fabrication) ou des résultats (de calcul) ;

- *règle* : ce mot contient la notion, très importante, d'obligation.

Il est bien certain que chacun des concepts apparus dans la définition du mot "logiciel" peut servir dans d'autres définitions à définir des objets qui n'ont rien à voir avec des logiciels. Autrement dit (si l'on utilise le vocabulaire de la théorie des ensembles), la définition du mot "logiciel" est une intersection de concepts et non une concaténation d'ensembles caractéristiques. Toute tentative qui consisterait à prendre chacun des concepts séparément, pour le démonter et essayer de voir en quoi il est caractéristique du concept de logiciel, est donc évidemment vaine.

En effet, chacun des concepts pris séparément couvre, pour une part éventuellement minime, la notion de logiciel mais surtout de très nombreuses autres notions qui n'ont rien à voir avec le logiciel. Il y a bien d'autres choses que des logiciels qui sont des programmes et il y a bien d'autres objets que des logiciels qui servent à faire fonctionner des ordinateurs. Nous expliquerons en détail plus loin, les raisons théoriques qui expliquent pourquoi structurellement une définition, dans n'importe quelle langue humaine, est nécessairement incomplète et ne peut absolument jamais être explicitée par la procédure qui consiste à remplacer chacun des termes qui la constitue par la définition de ce terme.

Ne pouvant procéder à ces remplacements et à cette recherche de traits caractéristiques, la seule question qui peut être

posée sur une définition comme celle du mot "logiciel " est celle de la pertinence du choix des mots qui la composent en se gardant d'aller trop loin dans l'explication de ces mots car on entre alors dans un processus de discussion infinie.

Pour conclure sur cette remarque préalable, constatons qu'elle amène à se poser la question suivante :

Puisqu'aucun mot du vocabulaire d'une langue humaine n'est défini de façon absolue et objective, mais uniquement par des références à d'autres mots, eux-mêmes non définis absolument, comment se fait-il que les hommes se comprennent ?

Cette question, très ancienne, est absolument fondamentale ; c'est elle qui, entre autres, entraîna la création de la linguistique. C'est également elle que cherchent à éclaircir ce que l'on appelle les neurosciences.

C'est dans un autre champ des sciences humaines, qu'une hypothèse de réponse a été fournie dès le début des années 1960.

Harold Garfinkel¹ propose de considérer qu'il n'y a pas de compréhension absolue, qu'il n'y a pas de sens absolu et que chaque être humain à tout instant de sa vie pratique, a le sentiment de comprendre ses congénères parce qu'il partage quelque chose avec eux (il partage des mots, il partage des sensations, il partage des perceptions, des interprétations...).

Cette hypothèse sur ce que pourrait être (ou plutôt, ne pas être) la compréhension explique en particulier pourquoi les

1.- GARFINKEL Harold, *Studies in ethnomethodology*, Prentice Hall Inc., New-York, U.S.A., 1967.

ordinateurs d'aujourd'hui ne comprennent pas ce qu'ils font (au sens humain du terme) et donne, pour l'avenir, des pistes de recherche de première importance pour fabriquer des machines qui pourraient partager avec les hommes suffisamment de choses pour qu'un sentiment de compréhension mutuelle s'installe.

3.1.2.2.- Qui traite des données ?

Dans la définition de la Commission nationale de terminologie, le mot "traitement de données" a été préféré au mot "ordinateur" pour des raisons que nous allons maintenant expliquer.

Il existe aujourd'hui déjà de très fortes présomptions sur le fait qu'il n'y a pas que des ordinateurs (tels que nous les connaissons aujourd'hui) qui fonctionnent en utilisant des logiciels. En utilisant le vocable "traitement de données", la définition française est simplement restée dans la logique que nous expliquons plus haut : elle a privilégié l'utilisation d'un terme un peu plus général au détriment du terme "ordinateur" jugé trop spécifique ; cela ne retirait rien à la précision de la définition puisque chacun des autres termes utilisés dans la définition couvre également des concepts étrangers à la notion de logiciel.

L'utilisation du vocable "traitement de données" permet également d'éviter que la définition vieillisse brutalement le jour où il apparaîtra évident à tout à chacun, qu'un génome, dans sa représentation codée par des humains, est également un logiciel. Nous nous garderons bien ici de tirer les conclusions juridiques qui pourraient découler de cette définition dans d'autres domaines que celui qui fait l'objet de

cette thèse. En particulier, la définition du logiciel qui a été élaborée par la Commission nationale de terminologie, semble permettre de couvrir de façon très efficace, le champ des séquences de gènes. On voit les conséquences d'une telle assimilation sur la protection éventuelle de ces séquences de gènes, qui pourraient alors éventuellement relever du droit d'auteur¹.

Remarquons incidemment que l'on a, pour les séquences de gènes, les mêmes craintes que l'on avait eu dans les années 1960 lorsque les logiciels avaient été exclus du droit à la brevetabilité. On craignait alors qu'en autorisant leur brevetabilité, on bloquerait le développement de l'informatique naissante parce que les entreprises dominantes de l'époque (en particulier les entreprises américaines) s'empresseraient de déposer des brevets. On pourrait assez facilement imaginer que, les mêmes craintes ayant les mêmes effets, les législations sur la protection des séquences génétiques artificielles ou sur la protection des découvreurs des séquences génétiques naturelles s'éloignent de la brevetabilité pour se rapprocher des textes sur le droit d'auteur.

3.1.3.- Qu'est-ce qu'un logiciel ?

Ayant éliminé la procédure d'interprétation la plus souvent utilisée pour analyser une définition, essayons maintenant d'examiner cette définition plus globalement. Notre examen bien entendu, gardera sa finalité juridique. Ainsi pour commencer, ne remettons pas en cause le caractère fonctionnel de la définition.

1.- Voir, à ce sujet, les travaux de Marie-Angèle Hermitte.

Remarquons tout d'abord que dans la définition n'apparaissent pas des mots qui permettraient de constater que le logiciel soit une œuvre de langage. Posons-nous donc pour commencer une question préalable :

- un logiciel peut-il être autre chose qu'une œuvre de langage ?

Un système de traitement de données est une machine. Peu importe sa technologie ; elle peut être en bois, en silicium ou en carbone. Le point important est qu'elle est distinguée des autres machines par le fait que sa matière première est de l'information et que sa production est également de l'information et surtout que le traitement qu'elle fait subir à la matière première se place dans le cadre d'un système logique qui s'exprime sous une forme langagière très particulière que nous allons maintenant expliquer.

On pourrait se demander, pour illustrer cette première idée s'il existe des machines produisant autre chose que de l'information tout en utilisant, comme entrée, des séquences d'informations variables. De telles machines existent évidemment ; ce sont par exemple, les orgues de Barbarie et les métiers à tisser mécaniques. Dans ces deux machines, il est possible de changer la forme, les couleurs sonores ou visuelles, en utilisant des bandes de carton perforé ou des rouleaux à picot. A chaque carton ou à chaque rouleau correspondra une musique ou un tissage différent.

Les cartons et les rouleaux sont-ils des logiciels ? La réponse à cette question est négative. Ce sont bien des programmes en ce sens qu'ils déterminent la séquence des couleurs et des sons produits par la machine. Ce ne sont pas des logiciels en ce sens que ce ne sont pas des séquences composées d'opérateurs

et d'opérandes qui décriraient l'exécution d'un calcul. Dit d'une autre façon et, pour utiliser la terminologie moderne, les bandes cartonnées et les rouleaux à picot sont des fichiers de données.

Mais où est donc le logiciel dans le métier à tisser ou l'orgue de Barbarie ? Il n'y en a pas, car la machine n'est pas un interpréteur de séquences d'opérateurs. La machine est un opérateur unique qui, de plus, ne réalise pas un calcul débouchant sur une nouvelle information, mais qui fabrique un objet physique.

3.1.3.1.- Codes de programmations et de commandes

Poursuivons maintenant en examinant deux objets imaginaires qui auraient pu exister techniquement dès la fin du dix-huitième siècle.

Ces objets seraient constitués par une machine à additionner de Pascal dans laquelle les données à additionner seraient, tout comme dans l'orgue de Barbarie, entrées dans la machine par l'intermédiaire d'un ruban cartonné. Dans la première de ces deux machines imaginaires, il n'y aurait sur le ruban que des nombres et la machine additionnerait les nombres aux nombres précédents. Dans la deuxième machine, il y aurait sur les rubans des séquences de nombres séparées par une configuration particulière de picots qui remettraient l'additionneur à zéro. La deuxième machine imaginaire additionnerait donc les nombres d'une série jusqu'à ce qu'elle rencontre la configuration particulière de remise à zéro, puis commencerait une nouvelle addition sur la série de nombres qui suivrait cette configuration particulière.

La bande de la première machine ne serait incontestablement pas un logiciel pour les mêmes raisons que le métier à tisser mécanique n'est pas un logiciel. On voit ici que la différence entre la première machine imaginaire et le métier à tisser réside dans le fait que la première machine imaginaire traite de l'information (elle ne produit pas du tissu, mais le résultat des additions), ce qui n'est pas le cas du métier à tisser. Cette différence n'est pourtant pas suffisante pour transformer un fichier de données en un logiciel.

Examinons maintenant les différences entre nos deux machines de Pascal imaginaires.

La bande de la seconde machine contient, sous la forme d'une configuration de perforation particulière, un objet qui n'est plus une donnée mais une commande à la machine elle-même. Il y a donc dans la mécanique de la seconde machine, un dispositif, activé par cette configuration particulière de perforations, qui interrompt le déroulement du fonctionnement normal de la machine pour lui faire commencer un nouveau processus. Ce système mécanique pourrait être absolument identique à celui qui permet dans une automobile de remettre à zéro la totalisation kilométrique journalière.

A la question de savoir si la bande perforée de notre seconde machine imaginaire constitue un logiciel, la réponse est, à notre avis, toujours négative. En effet, le "langage de commande" de cette machine est constitué d'un seul mot explicite : "l'ordre de remise à zéro". L'ordre d'addition n'est, pour sa part, pas explicite. Certes théoriquement un langage de programmation peut n'être constitué que d'un seul mot mais cela reste un cas d'école. Si notre machine de Pascal avait été capable de faire non seulement des additions mais aussi des soustractions et quelques autres opérations, alors très

clairement, la bande perforée qui servirait à la commander serait un logiciel. En effet, la bande contiendrait bien un programme, un procédé, une règle, destinés à faire fonctionner la machine, c'est-à-dire à lui transmettre des commandes en vue de lui faire exécuter un traitement dont la composition peut être infiniment variable et fait appel à des opérateurs de calcul multiples dont la juxtaposition des mécanismes physiques correspondants constitue la machine.

L'inventeur de cette machine serait obligé, pour en indiquer le fonctionnement, d'exhiber une table présentant, en vis à vis de chaque configuration particulière de perforations, l'effet de cette configuration sur la machine : remise à zéro, activation d'un sous-ensemble mécanique réalisant une addition, une soustraction ou tout autre opération. Dans cette table, chaque configuration particulière de perforations pourrait être désignée à l'aide d'un mnémonique évitant au programme humain d'avoir à faire l'effort désagréable de se souvenir de chaque configuration particulière.

On voit ici très clairement que nous sommes passés d'une mécanique à opérateur unique à une mécanique à opérateurs multiples et que cette transition s'est accompagnée de l'impérieuse nécessité de la définition d'un code : à chaque opération réalisable par la machine correspond une configuration de perforations, l'ensemble de ces configurations constituant le code de programmation et de commande de la machine. Ce code est déjà un langage. Les mnémoniques associés à chacune des configurations de ce code, s'ils représentent pour des humains des formes langagières plus accessibles se dérivent de façon complètement automatique du code. De même que l'on préfère désigner un livre par son titre plutôt que par son numéro d'ISBN, l'humain préfère manipuler le code mnémonique (le

langage formé par la liste des mnémoniques) plutôt que le code (les configurations de perforations).

Remarquons que ce n'est pas parce qu'il y a un code, qu'il y a un langage de commande. Si un code n'est pas toujours un langage, par contre un langage est toujours un code. Aussi, l'expression commune "langage codé" n'a pas grand sens. Il faut l'entendre comme servant à désigner des objets n'ayant pas l'apparence du langage naturel des humains. Ce terme est cependant beaucoup trop imprécis pour pouvoir être utilisé en tant que concept scientifique. Il ne faut donc pas s'étonner que les textes de juristes qui utilisent ce mot posent d'énormes problèmes d'interprétation.

Ainsi en est-il en particulier de la phrase du Sénateur Jolibois : "l'utilisation du logiciel passe par la lecture des instructions rédigées en un langage codé". L'imprécision de cette phrase permet à Jean-Michel Piot, par exemple, de dire qu'une partition musicale est également une œuvre de langage codé, ce qui est correct dans la langue commune mais tout aussi imprécis du point de vue des sciences du langage.

Les notes dans l'orgue de Barbarie sont également codées sous forme de perforations. De même, dans la seconde comme dans la première machine imaginaire, leur inventeur a nécessairement défini un codage, (différent du codage qui commande l'activation des opérateurs) qui associe à chaque chiffre de 0 à 9 ou à chaque nombre une configuration particulière de perforation. Ce qui permet de dire qu'une séquence de codes est un logiciel, c'est que parmi les codes qui figurent dans cette séquence, certains servent à désigner des opérations ou à donner des commandes à la machine.

Avant de poursuivre notre raisonnement, précisons un détail : la distinction faite par les informaticiens entre commandes et instructions.

La définition américaine du mot logiciel, contenue dans un amendement de 1980 à la loi de 1976 sur le copyright, dit : "*a computer programme is a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result*". Deux mots très importants de cette définition sont "*statements*" et "*instructions*". En français, ils ont respectivement pour équivalent le mot "commande" et le mot "instruction". La traduction de "*statements*" par "énoncé" est donc manifestement erronée et entraîne un trouble dans la compréhension de cette définition par le locuteur français.

Rappelons pour mémoire, qu'en français comme en américain, la différence entre une commande et une instruction données à un ordinateur, c'est que les commandes s'adressent au système d'exploitation et ont trait au fonctionnement même de la machine (mise en marche d'une imprimante, effacement d'un écran, recopie d'un fichier) alors que les instructions servent à définir les traitements qui sont appliqués aux données. Bien entendu, les commandes peuvent être incluses dans le corps d'un logiciel qui sera cependant composé essentiellement d'instructions.

Il existe toutefois des objets particuliers composés uniquement de séquences de commandes. Sur un ordinateur fonctionnant sous système MS-DOS, ce seront les séquences contenues dans les fichiers dont le nom se termine par ".BAT". Ces séquences de commandes servent à programmer entre autre le démarrage de l'ordinateur (affichage de l'heure et de la date, initialisation de la mémoire, chargement d'un logiciel, positionnement dans un répertoire de fichiers).

Un autre exemple d'objets composés uniquement de séquences de commandes (et des paramètres associés) est celui du fichier qui programme les mouvements d'un robot ou d'une machine outil à commande numérique.

A notre avis, ces objets particuliers doivent aussi être considérés comme des logiciels bien qu'ils ne réalisent pas, à proprement parler, des traitements de données. Ce choix est relativement arbitraire car on pourrait tout aussi bien considérer par exemple que la bande de pilotage d'un robot entretient avec l'objet qu'elle sert à fabriquer une relation identique à celle qu'entretient avec cet objet une photographie, une description technique. Bien que d'une forme différente de celle d'un plan ou d'une photographie, la bande de pilotage relèverait donc plutôt de la loi du 11 mars 1957 et non de la loi de 1985, au même titre qu'un logiciel classique.

Notre position s'explique cependant par le fait qu'il nous semble que beaucoup trop souvent, les juristes qui traitent de la question des logiciels ont une vision très restrictive de ce qu'est effectivement un logiciel.

Très grossièrement, on pourrait dire que pour un non informaticien, un logiciel c'est soit un logiciel de traitement de texte, soit un tableur, soit un SGBD (système de gestion de base de données), soit un système-expert. Ces quatre logiciels ont tous en commun de faire appel massivement à un utilisateur qui entre des données sur lesquelles sont appliqués des traitements.

En terme de nombre d'unités disséminées de par le monde, l'immense majorité des logiciels ne correspond à aucun de ces quatre genres. Ce sont les logiciels qui font fonctionner les

montres de poignet, les fours à micro-ondes, les chaînes stéréo, les machines outils et dans lesquels l'utilisateur n'introduit que peu ou pas de données, pas d'instructions et très peu de commandes qui sont les plus nombreux.

3.1.3.2.- Des codes aux langages

Nous avons vu que ce qui est important dans la définition du logiciel, c'est que la séquence de codes contient des instructions ou des commandes à la machine. Rappelons pour illustrer cette importance, quelques éléments de l'histoire de la naissance de l'informatique. Il n'est pas question ici d'entrer dans le détail d'une énonciation exhaustive. Ainsi ne mentionnerons-nous que trois noms : Boole, Babbage et Turing.

Dès 1847, l'anglais Georges Boole pose les bases d'un système formel qui permet de réaliser un grand nombre de calculs en restant strictement dans le cadre d'une logique binaire.

Vers 1940, un autre mathématicien, Charles Babbage définit une machine à calculer analytique dont on considère usuellement que c'est la première tentative de traitement automatique de l'information. En effet, cette machine (qui ne fut jamais construite) contenait explicitement la notion de programme, la notion de mémoire, la notion d'instruction et la notion de données. Vers 1940, un troisième mathématicien également anglais, Alan Mathison Turing, franchit une nouvelle étape en proposant un modèle abstrait d'automate à nombre d'états fini et en démontrant que toute fonction calculable peut être calculée par cet automate.

Les travaux de ces mathématiciens seraient restés purement théoriques si en 1948, n'avait pas été inventé le transistor. On savait, depuis une vingtaine d'années déjà, fabriquer des

machines électromécaniques dont la logique de fonctionnement permettrait de concrétiser les opérateurs booléens tout comme les roues dentées de la machine de Pascal concrétisait l'addition. Mais ces machines électromécaniques étaient d'une complexité et d'une fragilité qui leur interdisait de dépasser la puissance d'une simple calculatrice actuelle.

Dès le début des années 50, on dispose donc à la fois des éléments théoriques permettant de définir un ensemble cohérent d'opérateurs de calcul, des codages permettant de représenter les opérateurs et les données ainsi que les machines capables de lire des séquences de codes et d'exécuter les opérations demandées.

On remarquera que depuis le début de ce chapitre, nous avons utilisé le terme "langage" seulement pour décrire des objets très particuliers que sont les mnémoniques associés au code qui servent à représenter les commandes d'activation et les opérateurs de calcul de la machine.

Voici un exemple de mnémonique dans le cas de la deuxième machine imaginaire que nous évoquions plus haut : supposons que la configuration de perforations qui remet l'additionneur en position zéro soit la configuration où tous les trous possibles dans la largeur de la bande sont percés. C'est cette configuration et elle seule, qui provoquera la remise à zéro. Appelons la RAZ. Ces trois lettres forment le mnémonique qui évite à l'utilisateur de la machine d'avoir à se souvenir de la configuration des perforations.

Dans la machine qui comprendrait plusieurs opérateurs, à chaque opérateur serait associé dans la table des codages une configuration particulière de perforations. Par exemple, la configuration d'activation de l'additionneur serait constituée

d'une ligne de perforations complètes sauf en première position à partir du bord droit.

De même, une configuration d'activation de la mécanique de soustraction pourrait être constituée d'une ligne de perforation complète sauf la deuxième position à partir du bord droit. Enfin, en dehors de ces codes associés à chaque opération, un autre ensemble de dix configurations permettrait de représenter les nombres de 0 à 9.

Le "texte" réel du programme est constitué par la bande et ses perforations, chaque série de perforations représentant soit un opérateur, soit un nombre. L'humain qui rédige ce programme peut le faire avec une épingle sur une bande vierge. On imagine facilement cependant qu'il vaut mieux qu'il fasse une première rédaction sur papier en utilisant les mnémoniques des instructions (RAZ pour remise à zéro, ADD pour addition, SOU pour soustraire ainsi que les nombres de 0 à 9).

Ayant écrit le programme sous cette forme, il rédigera la version perforée du programme. Cette seconde rédaction peut être sous-traitée à un simple exécutant, inintelligent mais soigneux, car elle ne demande que des consultations de la table du correspondances qui constitue le code.

Rien n'interdit, bien entendu, d'utiliser ces mnémoniques de base pour définir, par composition, des opérations nouvelles. Ainsi la multiplication est composée d'une séquence d'additions et la division d'une séquence des soustractions.

On peut alors attribuer à ces compositions des mnémoniques qui désigneront non pas des configurations de perforations mais des séquences constituées uniquement de mnémoniques

de base. Nous venons de décrire ici le processus qui a permis de définir successivement :

- des ordinateurs programmés en binaire ;
- un langage de programmation structurellement attaché à la machine (assembleur) et qui est constitué des mnémoniques de chacune des opérations mécaniquement (ou électroniquement) exécutable ;
- et enfin, un langage de programmation, dit de haut niveau, dont les instructions sont des empilages d'instructions des langages de niveau inférieur.

Quelle que soit la sophistication des empilages et quelle que soit l'apparence des mnémoniques (et y compris s'ils ont l'apparence des mots du langage des humains), les langages de programmation ne sont finalement que des combinaisons de codes activant les opérateurs de calcul mécaniquement (ou électroniquement) existant dans la machine. Les traitements qui peuvent être exécutés par des logiciels écrits dans des langages de haut niveau sont donc strictement et exclusivement ceux qui sont prévus par la théorie qui décrit les opérateurs de base.

Les ordinateurs actuels sont construits sur des principes théoriques (la logique booléenne et la machine de Turing) qui n'ont absolument aucune raison d'être capables de représenter autre chose que ce pourquoi ils ont été construits. En l'état actuel des connaissances, les mécanismes du langage humain semblent dépasser très largement le cadre de la logique booléenne et peut être même celle de la théorie des automates à nombre d'états fini. Le débat, non tranché à ce jour, est le suivant :

- soit on considère l'humain comme un "être" échappant au moins partiellement à la démarche de la science à travers les notions d'esprit ou d'âme ;

- soit on réifie l'humain en le limitant à son enveloppe charnelle. Dans ce cas, il est constitué d'un nombre fini de molécules qui ne peuvent prendre qu'un nombre fini d'états. Ce nombre est actuellement trop grand pour pouvoir faire l'objet d'une modélisation complète. On peut espérer que le progrès des connaissances scientifiques le permettra. Dans cette hypothèse, il n'y aurait pas de différence conceptuelle fondamentale entre un humain et un ordinateur, si ce n'est le nombre d'états différents susceptibles d'être pris.

Les ordinateurs ne pourront donc manipuler le langage des humains qu'à l'intérieur de leur cadre de fonctionnement, ce qui leur interdit, jusqu'à plus ample informé, de réaliser la plus grande partie des opérations langagières communément réalisées par un humain.

Du point de vue que nous venons d'exposer, un logiciel pourrait être considéré plus comme une œuvre de codage que comme une œuvre de langage. En effet, l'écriture du logiciel pourrait, de façon extrêmement restrictive, être ramenée à une simple activité de codage d'un discours (en langage humain) dans le codage interprétable par la machine. Il est d'ailleurs intéressant de constater que les informaticiens désignent du nom de "codage" la partie de leur activité qui consiste à passer d'un organigramme (explicitant de façon totalement détaillée ce que la machine doit faire et donc résultant de la phase "intelligente" de l'activité de programmation) au texte du logiciel exprimé dans le langage de programmation. Comme nous le verrons plus loin, il nous semble que le logiciel doit plutôt être considéré comme le résultat de l'ensemble du

processus de programmation que comme le résultat de la dernière étape, désignée par les informaticiens avec le terme de codage, mais qui, en pratique, n'est pas un codage mécanique, comme celui que réalise, par exemple, un logiciel de compilation.

3.1.4. - Langages naturels et langages artificiels, les critères de distinction

Nous allons examiner maintenant de quelles catégories d'œuvres de langage relève un logiciel.

Faisons une partition brutale de l'ensemble des langages. Cette partition sera faite à partir du critère de sélection suivant :

- il y a d'une part les langages qui reposent sur un ensemble de postulats initiaux et dans lesquels tous les mots du langage sont définis par une combinaison de ces postulats initiaux (que l'on appellera les primitives du langage).
- il y a d'autre part, les langages dans lesquels chaque mot est défini au moyen de combinaisons d'autres mots du langage sans cependant qu'il soit possible d'identifier un petit nombre de mots qui, eux, constitueraient les primitives du langage.

La méthode pratique de classification de n'importe quel langage en utilisant ce critère est la méthode dite des remplacements.

Elle consiste à prendre la définition d'un mot et à remplacer, dans cette définition, chaque mot par sa définition. Dans tous les langages construits à partir de primitives, il arrivera nécessairement un moment où la définition ne sera plus composée que de primitives.

Dans un langage sans primitives, deux cas peuvent apparaître :

- soit, après un certain nombre de remplacements, le mot se trouve défini par lui-même (exemple : *métallier* = *activité professionnelle consistant à travailler le métal* et *métal* = *matière travaillée par le métallier*) ;
- soit la définition augmente de volume jusqu'à utiliser tous les mots du langage sans converger.

Dans ce qui suit, nous appellerons la première catégorie "les langages artificiels" et la deuxième catégorie "les langages naturels".

Il n'entre pas dans le cadre de cette thèse d'expliciter le détail des raisons théoriques de la pertinence de ces catégorisations. Ces raisons découlent directement d'un ensemble considérable de travaux sur les langages formels, dont l'exposé détaillé nécessiterait après définition d'un formalisme, de nous éloigner considérablement de notre sujet¹.

C'est à contre-cœur que nous utilisons les termes de "langage artificiel" et "langage naturel" pour désigner ces deux classes de langages. Ces deux vocables sont cependant ceux qui sont utilisés, en particulier par les juristes. Ils ne sont pas adéquats car l'opposition des qualificatifs "artificiel" et "naturel" ne correspond pas au critère classificatoire : en réalité, le critère classificatoire est d'une part qu'il y a des

1.- On pourra consulter, entre autres, :
La théorie des grammaires formelles, de Noam CHOMSKY,
Formal properties of grammars in R.D. LUCE (dir.) - "Handbook of Mathematical Psychology" , New-York 1963,
Notions sur les grammaires formelles, de Maurice GROSS et André LENTIN - 2ème Edit. Gauthier/Villars - Paris 1970.

langages avec sémantique et d'autre part des langages sans sémantique, où seules des manipulations syntaxiques et morphologiques sont utilisées.

Une explication claire et concise de ce critère classificatoire est donnée par Jacques Hébenstreit dans l'encyclopédie Universalis¹ en prenant pour exemple le langage des mathématiques :

"Si l'on considère l'ensemble des mathématiques, on constate que les objets que l'on y définit, n'ont pas d'existence en dehors de leur définition et que leurs seules propriétés sont celles qui découlent de ces définitions".

Dans ces conditions, faire une démonstration, c'est combiner entre elles un certain nombre de propriétés selon des règles préétablies. Ces dernières sont l'objet de la logique mathématique, qui vise à formaliser le raisonnement déductif, c'est-à-dire justement à ramener le raisonnement mathématique à une manipulation de symboles sans signification. C'est en ce sens qu'il faut interpréter la célèbre boutade du logicien Bertrand Russel selon laquelle "les mathématiques sont une science où l'on ne sait pas de quoi l'on parle ni si ce qu'on dit est vrai". La démarche des logiciens visant à ramener la totalité des mathématiques à des manipulations formelles de symboles relève typiquement de l'informatique, même si les logiciens n'en ont pas toujours été conscients, et conduit, de ce point de vue, à considérer les mathématiques comme un domaine où la totalité de la sémantique est réductible à la syntaxe, ce que Léon Brillouin énonçait sous la forme suivante : "le contenu informationnel des mathématiques est nul."

1.- à l'entrée Informatique et mathématiques - page 1143.

Le qualificatif "artificiel" donne à penser que tous les langages de cette classe ont été inventés par l'homme et le qualificatif "naturel", par opposition, donne à penser que tous les langages de cette classe résulteraient d'un état naturel. Cette opposition est évidemment stupide puisque tous les langages sont inventés par l'homme et qu'il n'y en a pas de plus artificiels que d'autres.

Toutefois, on comprend ces qualificatifs dans le sens que certaines inventions résultent d'un processus volontaire et maîtrisé de création, alors que d'autres résultent d'un processus involontaire et non maîtrisé (la partie volontaire et maîtrisée serait les langages artificiels, l'autre serait les langages naturels).

Hélas cette interprétation ne correspond pas non plus à la réalité, car il existe plusieurs langues que nous classerons dans la catégorie des langage naturels et qui résultent cependant d'un processus d'invention volontaire et maîtrisé. L'espéranto est l'exemple le plus connu de ces langues.

Par ailleurs, de nombreux langages artificiels pourront voir les mots qui les composent faire l'objet de définitions non seulement dans le langage artificiel lui même, mais également dans le langage naturel. Ainsi en est-il par exemple des langages de programmation. La définition dans le langage artificiel de chacune des instructions et commandes est contenue dans le logiciel interpréteur ou compilateur. C'est un segment de logiciel qui viendra remplacer le mnémonique avant toute exécution par la machine.

Dans ce qui suit, lorsque nous parlerons de procédure de remplacement d'un mot par sa définition, il s'agira toujours de la définition rédigée dans un langage de même nature que celui du mot remplacé.

Cette mise au point faite, nous pouvons maintenant utiliser les termes de "langage artificiel" et "langage naturel" pour essayer de préciser dans quelle classe il convient de placer les différents objets dont le statut juridique est sujet à discussion.

3.1.4.1.- Les œuvres de langage et le critère d'apparence

Nous appellerons "œuvre de langage" tout objet composé de caractères alphabétiques, d'idéogrammes et signes typographiques.

Un œuvre de langage peut aussi bien être écrite dans un langage naturel ou dans un langage artificiel, que dans une combinaison des deux (un manuel de programmation ou un livre de mathématiques).

Rien n'interdit que les mots d'un langage artificiel aient la même graphie que les mots d'un langage naturel. Le critère d'apparence qui est souvent utilisé par les non-spécialistes pour dire que l'on programme en langage naturel, que l'on interroge une base de données en langage naturel, que la base de connaissances d'un système-expert est écrite en langage naturel, est absolument inepte de point de vue des sciences du langage. Au contraire, le critère que nous proposons permet de réaliser une classification qui ensuite, facilite considérablement la qualification juridique.

Nous allons donc tout d'abord examiner différentes œuvres de langage pour les qualifier soit d'œuvres de langage artificiel, soit d'œuvres de langage naturel.

Toutes celles qui seront qualifiées d'œuvres du langage naturel relèveront de la Loi du 11 mars 1957.

Parmi celles qui seront qualifiées d'œuvres de langage artificiel, certaines pourront, de plus, être qualifiées de logiciel et, à ce titre, relèveront de la Loi du 3 juillet 1985. Les autres, tout comme les œuvres de langage naturel, relèveront de la Loi de 1957.

Certaines œuvres de langage posent des problèmes de classification. En effet, pour que notre procédure de remplacement puisse s'appliquer, des conditions minimales sont nécessaires. La principale de ces conditions est que, en plus de l'œuvre de langage en cours de classification, il faut que soient disponibles des définitions des mots et que ces définitions soient contenues dans l'œuvre de langage. Si les définitions ne sont pas explicitement disponibles, c'est-à-dire contenues dans l'œuvre elle-même, le classificateur peut être amené à faire des suppositions qui découlent de l'apparence du texte. La classification devient alors totalement aléatoire. Nous allons expliquer pourquoi sur trois exemples :

Exemple 1 :

Dans certains langages de programmation, (comme LISP), le texte du logiciel peut être composé d'un seul mot, ce mot faisant également partie des mots que l'on peut trouver dans un dictionnaire du langage naturel.

Exemple 2 :

Des poèmes minimalistes (comme les Aïku japonais) peuvent être composés d'un nombre très faible de mots, la concision de la forme correspondant à une très grande importance de l'ellipse et du non dit.

Exemple 3 :

Une liste de mots rangés par ordre alphabétique peut aussi bien, comme nous allons le voir plus loin, être un simple sous-ensemble extrait d'une langue naturelle qu'une description d'ouvrage écrit dans un langage artificiel, appelé "thésaurus".

Ces trois exemples peuvent sembler être des cas limites et on pourrait penser que dans la quasi totalité des cas pratiques, le simple examen de l'apparence de l'œuvre de langage permet de faire la classification même si l'on ne dispose pas des définitions des mots contenus dans l'œuvre. Il n'en est rien. L'ambiguïté est générale et donc le lecteur moyen qui n'est pas au fait de cette indécidabilité commettra sans cesse des erreurs de classification comme celle qui est couramment commise à propos de la classification des bases de connaissances contenues dans les systèmes experts.

3.1.4.2.- Les œuvres de langage et le critère de finalité

L'indécidabilité sur l'apparence découle du fait que des textes ayant strictement la même apparence peuvent aussi bien être des œuvres de langage artificiel que des œuvres de langage naturel. En effet, il existe des langages artificiels qui, de façon délibérée respectent des règles de syntaxe et utilisent des graphies qui existent également dans des langages naturels. Le texte suivant est dans ce cas :

additionne 3 et 5
imprime le résultat de l'addition

Ce texte peut aussi bien être une séquence d'instructions, respectant la syntaxe et les mnémoniques d'un langage de programmation, qu'un exercice dans un livre d'initiation au calcul à l'école primaire. Dans le premier cas on sera en présence d'une œuvre de langage artificiel et même d'un logiciel, dans le deuxième cas on sera en présence d'un texte en langage naturel.

Cet exemple montre, *a contrario*, à quel point il est absolument indispensable de définir un logiciel, non seulement par sa forme mais également par sa finalité. Ce n'est donc pas un hasard si toutes les définitions légales ou non légales du logiciel utilisent le critère de finalité. Si la définition des logiciels n'était pas donnée, en particulier par leur finalité (ils servent à faire fonctionner les machines de traitement de données), le juriste serait confronté à un problème terrible. Seul le critère de finalité en effet permet de dire qu'un texte quelconque est aussi le texte d'un programme. Cela ne veut pas dire évidemment que ce texte n'est pas un texte en langage naturel. En effet rien n'interdit que le texte d'un logiciel soit défini par son auteur comme un poème. Par contre, pour qu'un auteur puisse dire qu'un texte est un logiciel, il faut qu'il exhibe l'ensemble du langage utilisé pour écrire le texte ainsi qu'une machine techniquement crédible, c'est à dire capable de l'exécuter.

Nous avons vu il y a quelques lignes que pour pouvoir appliquer notre procédure de remplacement, il fallait disposer des définitions des mots contenus dans l'œuvre elle-même. Dans le cas où elles sont absentes, et ne pouvant faire des suppositions qui conduisent à un classement aléatoire, on est conduit à demander ces définitions à la personne qui revendique la paternité de l'œuvre. Cette demande est absolument équivalente à demander à l'auteur d'exposer la

finalité de l'œuvre. En effet, l'auteur d'une œuvre poétique refusera de donner le dictionnaire, l'auteur d'une œuvre scientifique et technique donnera un dictionnaire de spécialité ou renverra à un dictionnaire du langage naturel déjà existant et l'auteur ayant écrit dans un langage artificiel présentera les définitions dans le langage correspondant.

Si le dictionnaire n'est pas exhibé, l'œuvre peut immédiatement être classée comme œuvre de langage naturel. Les poètes savent bien en effet qu'il est impossible de faire un dictionnaire universel.

Si un dictionnaire est exhibé, il est alors très facile de déterminer si l'œuvre est écrite en langage naturel ou en langage artificiel, par application de la procédure de remplacement.

On pourrait penser que nous venons là d'exposer un artifice où en réalité ce n'est pas la classification qui classe mais l'auteur de l'œuvre. Ce n'est pas exact, car l'auteur ne peut pas changer le statut de l'œuvre que le classificateur a déjà entre les mains. Il est en effet extrêmement difficile, pour ne pas dire impossible pour l'auteur, dans des circonstances pratiques précises, de construire *a posteriori* un dictionnaire autre que celui qui lui a servi *a priori* à écrire l'œuvre.

3.2.- Une classification des œuvres de langage

Dans ce sous-chapitre, nous examinerons et classerons les œuvres de langage suivantes : le thésaurus et les requêtes à des S.G.B.D., les logiciels et les organigrammes, les bases de connaissances et les systèmes experts.

3.2.1.- les thésaurii

Avant de procéder à la classification, rappelons ce qu'est un thésaurus.

Un thésaurus est composé exclusivement de mots dont les graphies sont empruntées au langage naturel. Dans un thésaurus les mots n'ont pas de flexions. Chaque mot du thésaurus est appelé mot-vedette. Ce mot représente (au sens de "est le symbole") une liste de mots qui ne font pas partie du thésaurus mais qui font partie du langage naturel et qui sont liés généralement entre eux par une relation de proximité sémantique.

Par exemple dans un thésaurus en médecine, le mot HOPITAL désignera la classe des mots du langage naturel qui sont proches du mot hôpital. Ce sont entre autres les mots clinique, centre hospitalier, hospice, centre de soins, dispensaire...

Les mots du thésaurus sont, de plus, reliés entre eux par des relations, qui sont généralement hiérarchiques. Ainsi HOPITAL sera le mot père d'autres mots du thésaurus, qui seraient par exemple MATERNITE, SERVICE DE CHIRURGIE, MORGUE... Chacun de ces mots fils est, tout comme le mot père, le symbole représentant une liste de mots du langage naturel.

Le thésaurus est donc composé, d'une part, des mots-vedettes et, d'autre part, des différentes relations que l'on a le droit d'utiliser pour relier entre eux ces mots-vedettes.

Un thésaurus sert à décrire des livres, des articles de périodiques, des peintures, des photographies, des films, etc. Une description utilisant les mots d'un thésaurus aura obligatoirement la forme d'une liste de mots-vedettes. La syntaxe de cette description est généralement très simple : les mots sont donnés par ordre alphabétique décroissant et séparés par des virgules.

Dans un thésaurus, un mot peut-être défini à l'aide des autres mots du thésaurus qui sont liés à ce mot par l'une des relations autorisées. Si l'on utilise notre procédure de remplacement, on voit que tout mot dans le thésaurus peut être remplacé soit par les feuilles terminales¹ qui lui sont inférieures, soit par sa racine mais qu'en tout cas la convergence se fait vers les primitives : la racine si les définitions de chaque mot sont ascendantes (FER est un fils de METAL), les feuilles terminales si les définitions sont descendantes (un METAL peut être du FER, du CUIVRE, de l'ALUMINIUM).

Un thésaurus fait donc clairement partie des langages artificiels. Cependant, la finalité d'un thésaurus, pas plus que la finalité des descriptions qui utilisent les mots du thésaurus, n'est pas de commander un ordinateur mais de décrire des livres ou différentes œuvres de l'esprit. En conséquence, ni un thésaurus, ni une description d'ouvrage

1.- On appelle "feuilles terminales", les mots du thésaurus qui sont au plus bas niveau hiérarchique, c'est-à-dire ceux qui ne sont pas les pères d'autres mots.

utilisant les mots d'un thésaurus ne peuvent être considérés comme des logiciels.

3.2.2.- Les requêtes à des S.G.B.D.

Lorsqu'on interroge une base de données bibliographiques, par exemple la base Pascal sur le serveur Questel, on est amené à écrire la requête en combinant des opérateurs booléens et des mots qui peuvent être :

- des mots du langage naturel (les mots du titre ou d'un résumé de l'ouvrage, le nom de l'auteur ou les mots-clés lorsqu'il n'y a pas de thésaurus associé à la base de données).

- des mots du thésaurus (lorsqu'un thésaurus est utilisé pour indexer les documents rentrés dans la banque de donnée).

Constatons en passant que, pour un lecteur non averti, le champ mot-clé d'une base de données bibliographique est ambigu puisque les mots qui sont présents dans ce champ peuvent aussi bien être des descripteurs quelconques (formant un sous-ensemble du langage naturel) que des mots choisis dans un thésaurus. Rien dans l'apparence de ces mots ne permet de lever l'ambiguïté. Il faut pour cela savoir si les indexeurs de la base de données utilisent un thésaurus ou non. Pour une fois, cette ambiguïté est sans importance sur le plan juridique puisque un thésaurus n'est pas un logiciel (une expression composée à partir de mots d'un thésaurus ne peut en aucune manière commander une machine). C'est donc toujours la loi de 1957 qui s'appliquera pour protéger la liste des descripteurs (mots-clés extraits du langage naturel ou mots-clés extraits d'un thésaurus) des fiches bibliographiques.

Il n'en est pas de même cependant pour le texte des requêtes. En effet, une requête à une base de données répond très précisément à la définition d'un logiciel. Elle est formée de commandes et d'instructions servant à organiser un traitement de données. Examinons ce point plus en détail en prenant pour exemple le langage de requête du système de gestion de base de données (S.G.B.D.) Questel-Plus, qui fonctionne sur le serveur Questel, où est implantée la base de données bibliographiques Pascal.

Ce langage de requête autorise l'utilisation :

- d'opérateurs booléens qui s'appliquent à des expressions (qui, elles-mêmes peuvent contenir des opérateurs booléens), par exemple :

exp. **ET** exp.

exp. **OU** exp.

exp. **SAUF** exp.

NON exp.

- de commandes à l'ordinateur, par exemple :

•• **HI** qui permet d'avoir l'historique d'une interrogation.

•• **ST** qui termine une session d'interrogation.

•• **VI** qui permet de visualiser le résultat d'une question.

L'utilisation des opérateurs et des commandes est soumise à une syntaxe rigoureuse même si relativement souple : l'utilisation des parenthèses pour indiquer l'imbrication des opérateurs booléens est facultative ; mais en cas d'omission il y a une imbrication implicite qui est calculée par la machine.

Tout comme certaines expressions dans des langages de programmation d'ordinateurs peuvent contenir des mots du langage naturel, qui ne seront pas traités comme des instructions¹, les instructions et commandes du langage de requête peuvent contenir des chaînes de caractères dont la graphie appartient au langage naturel ou à un thésaurus.

Ainsi dans la requête *histoire ET France*, l'opérateur **ET**, qui active un mécanisme d'intersection logique dans la machine, est précédé et suivi de deux paramètres qui sont les chaînes de caractères permettant de sélectionner les ensembles de fiches sur lesquelles l'opérateur va travailler. Bien entendu seules les instructions et les commandes dans un langage de programmation, comme dans un langage de requête, font partie du langage.

En prenant n'importe quelle requête syntaxiquement correcte, on peut procéder au remplacement des instructions et des commandes par leur définition c'est-à-dire par les logiciels qu'ils activent et, de proche en proche, on aboutira ainsi aux primitives de la machine sur laquelle est implanté le S.G.B.D.

Une requête à une base de données bibliographiques est donc bien une phrase d'un langage artificiel et, de plus, on doit la considérer comme un logiciel, puisqu'elle sert bien à programmer une série d'opérations dans la machine.

Afin d'être tout à fait complet, il nous semble nécessaire d'évoquer maintenant le cas de certains systèmes documentaires de conception récente (moins de quinze ans).

1.- l'instruction PRINT "bonjour" en BASIC contient une instruction PRINT et le mot bonjour, qui n'a pas de sens pour la machine et qui va être imprimé.

En France, deux exemples de tels systèmes peuvent être donnés : il s'agit du système SPIRIT de la société SYSTEX et du système DARWIN de la société CORA. Les bases de données qui peuvent être traitées par ces S.G.B.D. sont absolument quelconques (y compris des bases de données bibliographiques et des banques de données en texte intégral). Lorsqu'un nouveau document est entré dans la base de données, c'est le logiciel qui procède à son indexation automatiquement¹.

Les mots-clés résultant de cette indexation ne sont pas intégrés aux documents contenus dans les bases de données mais sont cependant accessibles à l'utilisateur. Les méthodes d'extraction de mots-clés peuvent être linguistiques (par exemple recherche des groupes nominaux dans le texte des documents entrés) ou statistiques (par exemple recherche des mots qui sont fréquents dans le document entré mais rares dans la langue générale).

Les requêtes d'un usager à de tels systèmes documentaires n'ont pas besoin de respecter une syntaxe artificielle. En effet, le système analyse la requête de la même façon qu'il a analysé les documents entrés dans la base de données. Puis il compare les mots-clés extraits de la requête avec les mots-clés extraits de chacun des documents contenus dans la base des données et il ramène, en réponse à la requête, les documents dont les mots-clés sont les plus proches (statistiquement ou linguistiquement) des mots-clés de la requête.

On voit donc que, dans ces systèmes, on peut tout à fait utiliser comme requête un segment de texte qui traite du sujet recherché. Ainsi, sur une base de données juridiques implantée

1.- Alors que dans les bases de données comme Pascal, ce sont des documentalistes humains qui ont choisi, éventuellement dans un thésaurus, les mots-clés qui servent à indexer les ouvrages.

au Conseil d'Etat sous le logiciel SPIRIT, on recherche un texte ou un arrêt en donnant comme requête d'autres morceaux de texte ou d'arrêts.

On peut alors se demander, puisque la requête est intégralement écrite en langage naturel, si celle-ci ne peut pas être aussi considérée comme un logiciel. Un segment de texte, dans le cadre de son utilisation comme requête à un S.G.B.D., tel SPIRIT ou DARWIN, répond en effet très précisément à la définition d'un langage artificiel d'une part, et d'un logiciel d'autre part.

C'est un langage artificiel très particulier en ce sens que la requête ne contient ni commande, ni instruction mais que, implicitement, la machine considère l'entrée de la requête comme un ordre d'activation de son logiciel de recherche ; la machine introduit des opérateurs logiques entre les mots-clés extraits du texte de cette requête. Dans SPIRIT et dans DARWIN, c'est le logiciel qui construit automatiquement, à partir du segment de texte en langage naturel fourni par l'utilisateur, la véritable requête, (c'est-à-dire le véritable logiciel) qui, dans le cas de Questel, doit être rédigée par l'utilisateur lui-même. Des logiciels tout à fait comparables à DARWIN et SPIRIT sont commercialisés pour venir s'interposer entre des usagers et des S.G.B.D. déjà existants mais dont le langage de requête (par exemple S.Q.L.) n'est pas facile à manipuler.

Notre position est de considérer qu'un segment de texte en langage naturel ne peut-être considéré comme un logiciel que dans le cas où, explicitement, sa finalité en tant que requête à un S.G.B.D. de type DARWIN ou SPIRIT, ferait partie du cadre de l'affaire en cours de jugement. Dans tous les autres cas un segment de texte initialement écrit en langage naturel

conserve son statut d'origine et n'est donc pas susceptible d'être protégé par la loi de 1985.

3.2.3.- Logiciels et organigrammes

Examinons maintenant des objets que les informaticiens désignent habituellement sous le nom de logiciels. Ensuite, nous examinerons des objets que les informaticiens désignent habituellement sous le nom d'organigrammes. Ayant effectué la classification de ces objets comme expressions de langage naturel ou de langage artificiel, nous examinerons s'ils peuvent être considérés comme des logiciels au sens de la définition qu'en donne la Commission nationale de terminologie de l'informatique.

Lorsque le texte d'un logiciel, qu'il soit rédigé en code binaire, en assembleur, en Fortran ou dans un langage de programmation semi-graphique (comme APL où, en plus des caractères alphabétiques latins, on utilise des caractères alphabétiques cyrilliques) est présenté au classificateur, l'ambiguïté fondamentale sur l'apparence que nous avons évoquée plus haut, est évidente. Ainsi, on a souvent vu, affichés sur les murs des salles d'ordinateurs, des pages couvertes de 0 et de 1 qui ne constituent pas le texte d'un logiciel. Ce sont des images numérisées (le portrait de Marilyn Monroe, le singe qui se gratte la tête, un Christ) qui ne se révèlent en tant qu'images que lorsqu'on est à une certaine distance de la feuille de papier.

3.2.3.1.- Compilateurs, interpréteurs, langages de commande

Il est rare qu'un logiciel contienne dans le corps même de son texte, les définitions des mots qui le composent. Ce sera

cependant le cas des logiciels qui servent à définir un langage de programmation dont tous les mots du langage sauf les primitives, sont décrits dans le langage lui-même. Ces logiciels particuliers font partie de la famille des compilateurs ou des interpréteurs de langage de programmation.

Dans le cas le plus général, un logiciel n'est cependant pas un compilateur ou un interpréteur de langage de programmation. Malgré tout, de très nombreux logiciels contiennent un segment qui est un interpréteur de langage artificiel. Ce langage est le langage des commandes du logiciel lui-même. Ainsi, dans le logiciel de traitement de texte Word, il y a un module interpréteur dont l'objet est d'exécuter et d'interpréter les commandes que l'utilisateur donne à Word (LIT, IMPRIME UN TEXTE, QUITTE....) en les distinguant des textes qui pourraient être saisis.

On comprend l'utilité pour le concepteur de Word de disposer des touches de fonction et de la souris.... car cela lui permet de considérablement simplifier son module interpréteur en lui évitant d'avoir à faire une analyse linguistique fine des textes tapés au clavier.

Bien entendu, ces langages de commande ne sont pas tout à fait équivalents à des langages de programmation en ce sens qu'ils n'ont pas une vocation à décrire des logiciels généraux mais seulement à piloter et programmer la machine dans le cadre très restreint du logiciel dont ils font partie. On trouvera ainsi des langages de commande de traitement de texte, de tableur, de SGBD (les SGBD contiennent parfois deux langages de commandes différents : le langage qui sert à définir les bases de données et le langage qui sert à écrire les requêtes à une

base de données existante, comme le langage SQL), de machine-outil à commande numérique, de serveur vidéotex.

Dans certains cas, le langage de commande est suffisamment puissant (c'est-à-dire qu'il contient une suffisamment grande variété de commandes et d'instructions) pour pouvoir être utilisé comme un véritable langage de programmation pour écrire des applications qui dépassent largement l'objet initial du logiciel à l'occasion duquel ils ont été définis.

Ainsi le logiciel de gestion de base de données DBase contient un langage de commande suffisamment puissant pour qu'il soit possible, dans ce langage, d'écrire des applications typiques de bases de données (gestion d'un fichier d'adresses, gestion d'un stock de pièces, gestion d'un petit fichier bibliographique) mais également d'écrire une application de synthèse de musique...

Etant donné qu'il n'y a pas de réelle différence entre un langage de commande et un langage de programmation, on comprend que nous placions les textes écrits avec l'un quelconque de ces langages, dans la même catégorie, à savoir celle des logiciels.

3.2.3.2.- Les logiciels sont exprimés dans des langages artificiels

Revenons maintenant au problème de la distinction entre langage artificiel et langage naturel.

Nous avons vu que la seule façon de lever l'ambiguïté fondamentale, permettant de classer le texte de ce qui semble être un logiciel, est de demander à son auteur d'exhiber le dictionnaire du langage dans lequel a été écrit le texte.

Dans la totalité des cas où le texte est un logiciel, son auteur exhibera instantanément et sans problème le manuel de programmation qui a servi à écrire le logiciel ainsi que le logiciel interpréteur ou compilateur. Ce dernier permet à la machine de procéder au remplacement automatique des mnémoniques du langage de programmation utilisé par des mnémoniques d'un niveau de langage inférieur ou par les codes binaires des opérations de calcul électroniquement exécutables par le matériel informatique. Si l'auteur peut faire l'exhibition, alors le texte peut être classé dans la catégorie des langages artificiels.

Examinons maintenant si, en plus d'être un langage artificiel, un tel texte peut être considéré comme un logiciel au sens de la définition donnée par la Commission de terminologie.

Remarquons que l'œuvre de langage en cours de classification peut contenir des mots, voire des phrases, dont l'apparence est celle de mots ou de phrases du langage naturel.

Ainsi en est-il des deux exemples suivants :

Exemple 1 :

```
10 PRINT "fable"  
20 PRINT "Maître corbeau sur un arbre perché"  
30 PRINT "tenait en son bec un fromage"  
40 PRINT "....."
```

Exemple 2 :

```
10 REM Multiplicateur  
20 INPUT "donnez l'opérande"; A: REM saisie de l'opérande  
30 INPUT "donnez le facteur multiplicateur"; B: REM saisie du multiplicateur  
40 C = A * B : REM calcul
```

**Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques**

```
50 PRINT "le résultat est : "; C : REM affichage du résultat
60 INPUT "pour continuer, taper ENCORE"; D$ : REM attente de lecture
70 IF D$ ≠ "ENCORE" THEN 60 : REM gestion erreur de saisie
80 CLS : REM nettoyage de l'écran
90 GOTO 10 : REM et on recommence
```

Les mots et les phrases qui apparaissent dans ces deux exemples ne font pas tous partie du langage de programmation proprement dit. Cela apparaît clairement dans le fait qu'ils ne sont pas définis dans le manuel de programmation.

Dans le texte d'un logiciel, l'occurrence d'un mot qui n'est pas défini dans le manuel de programmation ou dans le logiciel lui-même, est totalement et strictement contrôlée, faute de quoi l'exécution du logiciel par la machine est impossible.

Dans notre premier exemple, le texte de la fable est, du point de vue informatique, considéré comme un paramètre non interprétable de l'instruction "PRINT" qui, dans le langage de programmation BASIC, donne un ordre d'affichage sur l'écran. L'interdiction d'évaluation du paramètre par la machine est indiquée par la présence du guillemet ouvrant après la séquence de lettres PRINT. Dans notre premier exemple, nous sommes donc typiquement en face d'un cas où il est possible de décomposer le texte en cours de classification.

Une décomposition possible ramènerait :

d'une part, un sous-ensemble composé uniquement de commandes et instructions contenues dans le manuel de programmation. Ce sous-ensemble pourrait être considéré comme du langage artificiel "pur" ;

- d'autre part, une fable qui existait bien antérieurement à l'invention des ordinateurs, c'est-à-dire une œuvre en langage naturel qui, sans contestation possible, relève exclusivement de la loi du 11 Mars 1957.

Mais rappelons que la classification que nous cherchons à réaliser ne porte pas sur chacun des deux éléments constitutifs de l'œuvre parce qu'alors le nombre de décompositions possibles est très grand. Ainsi, rien n'interdit de décomposer à son tour le sous-ensemble formé uniquement de commandes et d'instructions. On ne trouvera alors que des mots du langage naturel (les mots anglais PRINT, INPUT, etc.) ou, à la limite, que des lettres alphabétiques.

On voit bien sur cet exemple l'arbitraire complet, et donc l'irrationalité, des tentatives d'attribution de statut juridique à des œuvres de langage par des méthodes de décomposition. En effet, dès qu'une décomposition est amorcée, elle échappe à tout contrôle, puisque suivant les critères de segmentation choisis, on pourra conclure dans un sens ou dans un autre. Il ne faut donc pas l'engager et évaluer l'œuvre dans sa totalité en exécutant très strictement la procédure de remplacement.

Si nous appliquons très strictement à notre premier exemple les remplacements prévus par l'interpréteur du langage Basic, qui aura été fourni par l'auteur de ce texte, alors ce texte va être remplacé par un autre dans lequel l'instruction PRINT sera remplacée par le segment de logiciel qui commande à la machine l'affichage sur l'écran. Chacun des caractères des mots de la fable sera remplacé également par un code hexadécimal puis binaire. De proche en proche, le texte initial va être remplacé par un texte composé uniquement de 0 et de 1, directement exécutable par la machine, c'est à dire que certaines configurations binaires activeront des opérateurs

électroniques cependant que d'autres seront les données sur lesquelles ces opérateurs s'appliqueront.

Ce texte est donc bien une œuvre de langage artificiel, parce que nous avons pu remonter jusqu'aux primitives du langage et que ces primitives n'ont aucun sens autre que celui qui leur a été attribué par leurs concepteurs (à savoir activer les opérateurs de la machine).

Par ailleurs ce texte est un logiciel car il a bien pour finalité de commander un traitement de données.

3.2.3.3.- Machines et œuvres de langage

Nous venons de voir qu'il ne fallait pas utiliser une procédure de décomposition analytique pour qualifier une œuvre de langage artificiel puis de logiciel. On pourrait penser que, dans certains cas, il n'est pas possible de qualifier un objet sans pratiquer une décomposition. Nous n'avons pas trouvé de tels exemples. Bien au contraire, la procédure de remplacement que nous proposons, ainsi que la mise en œuvre de l'analyse globale, permettent de qualifier une gamme très large d'objets.

Ainsi, un robot qui contient un logiciel, un four à micro-onde, une montre à cristaux liquides, un lecteur de disques compacts ne peuvent en aucun cas être qualifiés de logiciels pour deux raisons, chacune d'entre elle étant éliminatoire :

- la première est que ce ne sont pas des œuvres de langage.
- la seconde est que ces objets ne sont pas "relatifs au fonctionnement d'un ensemble de traitement de données", c'est-à-dire ne servent pas à faire marcher, par exemple, un

ordinateur. Leur finalité est toute autre. Remarquons en particulier qu'un lecteur de disques compacts, même s'il réalise en interne, des traitements sophistiqués de données binaires (lecture et décodage, détection d'erreurs), n'a pour finalité que de produire de la musique et des sons et qu'il ne faut surtout pas engager le débat sur les composantes de cet appareil. Si on l'engage pour le lecteur de disques compacts, on doit aussi l'engager pour tout autre objet ; une chaise peut être considérée comme un système de traitement de données puisque détectant la force exercée par le postérieur qui s'assied sur elle, elle transmet cette force au sol et, en retour, transmet au postérieur, outre la force, une information qui est que le sol ne s'enfoncera pas. Bien que le traitement de données ne soit pas électronique, la chaise de ce point de vue, a très certainement traité des données.

3.2.3.4.- Les manuels de programmation, les listages

Demandons nous également si un manuel de programmation qui contient des textes de programmes est un logiciel.

Notre méthode classificatoire nous permet de dire que ce livre n'est pas un logiciel. C'est une œuvre de langage qui, lors de l'application de la procédure de remplacement, apparaîtra comme étant du langage naturel puisque, dès la première phrase de l'introduction de l'ouvrage, la procédure de remplacement se révélera être une tâche infinie.

Le seul cas où un ouvrage contenant des textes de logiciels est lui-même un logiciel, du point de vue de notre procédure de remplacement, est celui où l'ouvrage répond à deux contraintes.

- la première est que, du premier au dernier mot, on ne trouve dans l'ouvrage que des expressions d'un langage artificiel

(avec la possibilité bien entendu d'inclusions en langage naturel à l'intérieur des phrases en langage artificiel).

- la deuxième est que le langage artificiel qui est utilisé dans cet ouvrage, doit, au cours de la procédure de remplacement, déboucher sur des codages destinés à un ensemble de traitement de données.

Dit d'une autre façon ce livre très particulier est un listage de logiciel qui aurait fait l'objet d'une édition. La deuxième condition est indispensable pour distinguer le listage édité d'une autre édition en langage artificiel, qui ne serait pas du logiciel mais, par exemple, un thésaurus.

Du point de vue juridique, le texte du premier exemple (celui de la fable) est, sans contestation possible, protégeable par la loi du 3 juillet 1985. Si Jean De La Fontaine était vivant, une question intéressante serait de savoir si cette œuvre serait une œuvre dérivée de la sienne. De manière incontestable, cette œuvre logicielle n'est pas une copie servile, puisqu'elle contient de nombreux éléments supplémentaires que Monsieur De La Fontaine ne pouvait connaître à son époque et donc ne pouvait intégrer à sa fable. Il nous semble assez clair cependant que ce logiciel, dans la forme exacte où il est ici présenté, est une œuvre dérivée puisqu'elle ne pourrait exister si Monsieur De La Fontaine n'avait au préalable écrit sa fable. Le logiciel de notre exemple entretient avec la fable de Jean De La Fontaine un rapport analogue à celui qu'entretiendrait avec cette fable une traduction dans une autre langue, une adaptation cinématographique ou un opéra.

Remarquons, pour conclure les commentaires sur ce premier exemple, que la procédure de décomposition analytique (dont nous avons démontré l'ineptie lorsqu'il s'agissait de

déterminer le statut de l'œuvre dans son ensemble) devient parfaitement utilisable lorsqu'il s'agit d'essayer d'identifier à l'intérieur de l'œuvre, et comme élément constitutif, la forme d'autres œuvres (ici une fable, mais cela aurait pu être un autre logiciel, un graphisme, des sons...) dont elle serait dérivée.

Remarquons également que ces éventuelles œuvres originales sont difficiles à mettre en évidence dans certaines représentations du texte, alors qu'elle sont évidentes dans d'autres représentations du texte qui correspondent à d'autres étapes dans la procédure de remplacement. Ainsi la fable de Jean De La Fontaine apparaîtrait de façon beaucoup moins évidente (bien que tout aussi incontestable) dans la version binaire du logiciel de l'exemple n° 1.

Remarquons pour terminer que l'auteur de ce logiciel ne pourrait pas l'utiliser commercialement, sans l'autorisation de l'auteur de la fable (si celle ci n'était pas tombée dans le domaine public), car la moindre exécution du programme ferait apparaître sur l'écran de l'ordinateur ou sur un listing une copie servile de la fable. Cette copie servile serait clairement une contrefaçon car la seule différence avec l'œuvre originale serait le support et éventuellement la graphie.

3.2.3.5.- le *look and feel*

Abordons maintenant l'exemple n° 2. Nous ne référons pas la démonstration tout à fait analogue du fait que cette œuvre est également un logiciel.

La question qui se pose est donc uniquement de vérifier si l'œuvre est "absolument" originale (c'est-à-dire à la fois

originale et nouvelle) ou si elle dérive d'une œuvre antérieure. Pour ce faire, commençons par examiner successivement tous les mots qui ne sont pas définis dans le dictionnaire du langage artificiel :

- pour des raisons techniques, les noms des variables, qui en Basic sont limités à huit caractères, peuvent être très proches dans deux logiciels différents réalisant les mêmes fonctions. L'identité complète entre les noms des variables dans deux logiciels différents, pourvu qu'il soient assez longs, résulterait cependant d'un hasard très improbable. Cette identité, tout comme l'identité de la totalité des entrées dans deux dictionnaires de langue courante, pourrait donc à juste titre être considérée comme une preuve de dérivation, voire de copie servile.

- pour des raisons techniques, les commentaires, c'est-à-dire les segments des phrases qui suivent immédiatement l'instruction définissant un commentaire (en Basic c'est l'instruction REM), peuvent être très proches d'un logiciel à l'autre. Tout comme pour les noms des variables, l'identité partielle ou totale donnera une forte présomption de dérivation ou de copie servile.

- les messages affichés à l'écran ou imprimés, ainsi que les textes acceptés en entrée par le logiciel (qui pour une part importante constituent des commandes du logiciel en train de fonctionner par son utilisateur), forment la troisième et la dernière catégorie d'objets qui peuvent apparaître dans le texte du logiciel, sans pour autant faire partie du langage artificiel dans lequel est écrit le logiciel. Contrairement aux variables et aux commentaires, les messages et les commandes sont accessibles à l'utilisateur pendant l'exécution du logiciel. Ce sont eux en effet qui constituent ce que l'on

appelle usuellement l'interface homme/machine. Nous n'utiliserons cependant pas ce vocable qui est très imprécis car il peut également recouvrir des séquences d'instructions qui interprètent les commandes entrées par l'utilisateur ou qui impriment et affichent les messages émis par la machine.¹ C'est sur ces messages et sur les langages de commande que porte une part importante du débat juridique désigné par le vocable "*look and feel*".

En pratique le débat sur le *look and feel* porte sur l'ensemble interface homme/machine, c'est-à-dire inclut également le déroulement des séquences de communication entre l'homme et la machine. A cause de l'imprécision de ce terme, nous éviterons également de l'utiliser.

Tout comme ressemblance entre deux œuvres de langage naturel dépend de l'appréciation du juge, la similitude entre les messages et les langages de commande de deux logiciels sera évaluée au cas par cas.

Remarquons qu'une très grande proximité, voire une identité quasi-complète n'est pas, le plus souvent, estimée suffisante (en dehors de la présence de tout autre élément) pour que soit jugé qu'il y a dérivation ou copie. Ainsi le logiciel Foxbase reprend l'intégralité du langage de commande du logiciel DBase qui lui est antérieur ainsi que la plupart des messages de la machine à l'utilisateur. Le reste du logiciel, c'est-à-dire les instructions elles-mêmes ainsi que l'organisation interne du logiciel, sont clairement différents du logiciel antérieur. De plus le logiciel Foxbase contient des commandes qui n'existaient pas dans le logiciel DBase.

1.- Le mot "interface homme/machine" peut également couvrir des éléments beaucoup plus matériels, comme une électronique de synthèse de la parole ou un clavier.

Foxbase est typique de ce que l'on appelle un clone logiciel. En France, comme aux U.S.A., Foxbase n'est pas considéré comme dérivé de DBase. Il peut être en effet considéré que l'auteur du logiciel antérieur, en rendant publics les messages à l'utilisateur et le langage de commande de son logiciel (qui sont exprimés avec des mots du vocabulaire courant et dans une composition qui a pour objectif primordial l'évidence d'utilisation), ne peut pas revendiquer une protection, car cela conduirait à un blocage généralisé de l'activité de développement de nouveaux logiciels.

Tout comme les logiciels n'ont pas été protégés par les brevets, pour des raisons qui dépassent les intérêts d'éventuels déposants, le juge considère généralement que même une très forte proximité entre messages et langages de commandes de deux logiciels différents n'est pas suffisante pour donner au logiciel antérieur une part de propriété sur le logiciel postérieur.

Comme le soulignent justement les auteurs du Lamy-Droit de l'informatique :

- soit cette proximité se traduit par un emprunt de forme ; la contrefaçon pourra alors être établie si la création copiée est originale.
- soit "la sensation d'identité ou au moins de proximité entre les deux produits ne passe pas par un démarquage de la forme"¹ ; le juge ne pourra alors constater la contrefaçon.

1.- Lamy droit de l'informatique 1991, n° 1354.

3.2.3.6.- Les organigrammes

Après avoir examiné les éléments qui, dans un logiciel, ne sont pas du langage artificiel, étudions maintenant les séquences d'instruction d'un logiciel proprement dites.

Un informaticien considérera que si les séquences d'instructions d'un logiciel sont identiques à celles d'un autre logiciel (et même si toutes les variables portent des noms différents, même si les commentaires sont différents et même si le langage de commande et les messages à l'utilisateur sont différents), alors les deux logiciels seront identiques. Un informaticien expliquera sa position par le fait que le changement des noms des variables et le changement des commentaires n'ont strictement aucune influence sur l'apparence, pour l'usager, de l'exécution du logiciel par la machine.

Par ailleurs, les messages affichés, engendrés par la machine, ainsi que le langage de commande peuvent être complètement différents du point de vue formel, tout en gardant le même sens. Ainsi, la commande "IMPRIME" peut être synonyme de la commande "ECRIT", c'est-à-dire que l'interprétation du mnémonique (ou encore sa signification dans le langage naturel) est restée constante alors que sa forme graphique a changé.

Cette méthode de copie servile est, du point de vue technique, aisément démontrable. Il est cependant beaucoup plus délicat d'identifier les segments qui, dans un logiciel, ont été empruntés à un autre logiciel. Il y a trois raisons à cela :

- le langage de commande et les messages à l'utilisateur sont difficilement protégeables,

- les noms des variables et le texte des commentaires sont sans importance, car extrêmement faciles à modifier
- les mots qui restent lorsqu'on a retiré les précédents sont exclusivement des mots du langage de programmation, obligatoirement identiques dans deux logiciels qui ont été écrits dans le même langage,

La comparaison ne peut donc, exclusivement, porter que sur la composition. Mais la composition de deux logiciels ayant le même langage de commande et les mêmes messages à l'utilisateur (ce qui revient à dire qu'ils ont des fonctionnalités identiques ou très proches, sans que cela constitue un élément solide de protection pour le logiciel antérieur) peut-être nécessairement proche. Cependant, il peut exister plusieurs procédés, plusieurs méthodes, plusieurs algorithmes qui amènent au même résultat. Les méthodes et les procédés étant dans le domaine public, la comparaison sur les compositions de deux logiciels écrits dans le même langage, réalisant les mêmes fonctions et reposant sur le même algorithme, portera donc sur la mise en forme de cet algorithme.

L'étape intermédiaire entre l'algorithme (qui relève du domaine public) et le logiciel exécutable par la machine (explicitement protégé par la loi de 1985) est l'organigramme. Un organigramme est une œuvre de langage dont la composition, en particulier visuelle, est certes un peu particulière, mais qui néanmoins est incontestablement composée de mots et de symboles.

Cette façon de décrire un organigramme est absolument cohérente avec l'ensemble des travaux de la linguistique et de la sémiologie. En effet, ces sciences considèrent que l'on ne peut réduire les œuvres de langage à celles qui seraient

exprimées en caractères alphabétiques (latins, cyrilliques...). Sont aussi œuvres de langage celles qui sont exprimées en caractère idéographiques. Des caractères idéographiques sont utilisés non seulement pour écrire des langues naturelles (comme la langue chinoise) mais aussi pour écrire des langues artificielles (comme la langue des panneaux routiers)¹.

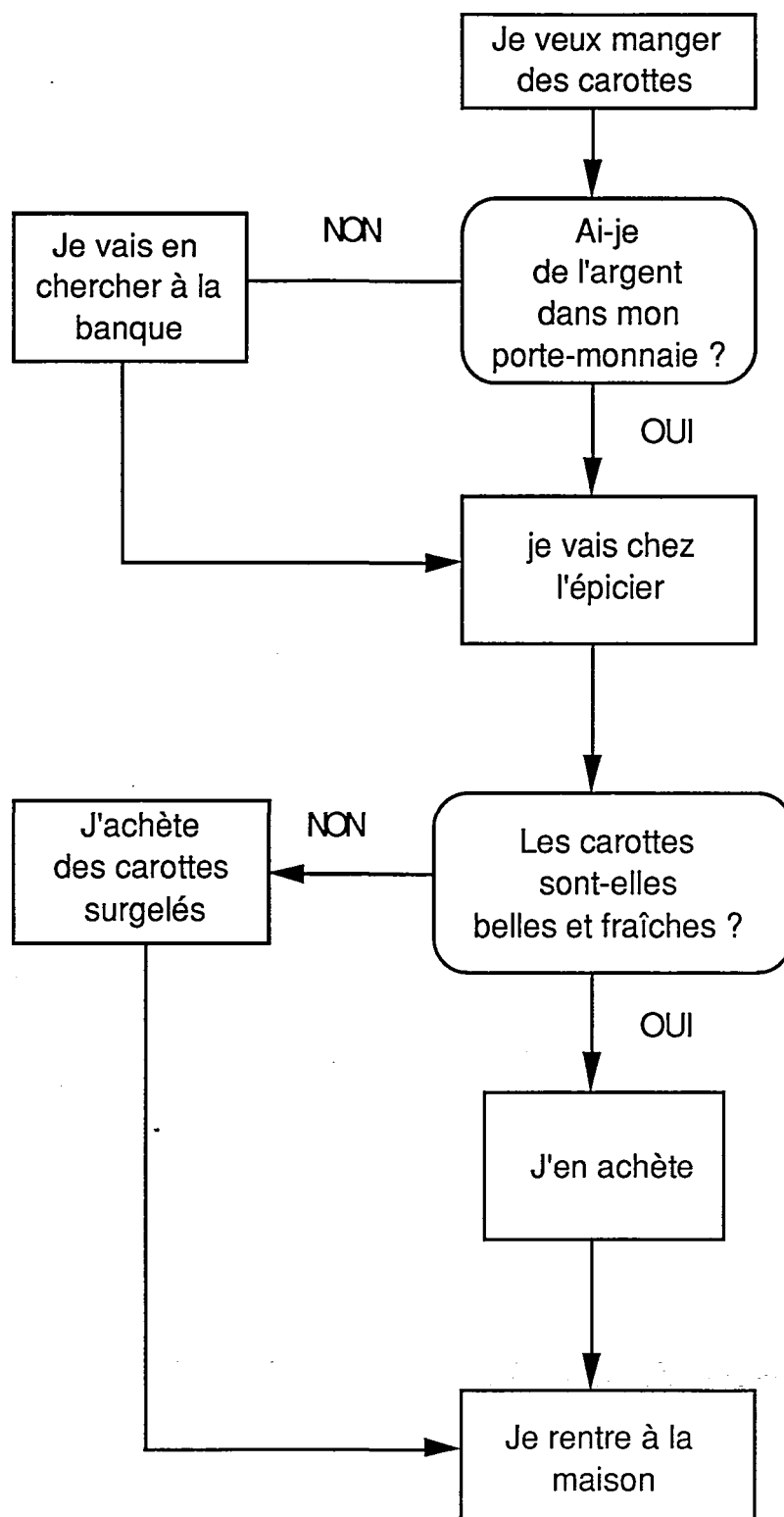
Il convient alors de garder à l'esprit la distinction entre, d'une part, les œuvres de langage idéographiques (naturelles et artificielles) pour lesquelles des dictionnaires peuvent-être exhibés et, d'autre part, les œuvres purement graphiques (tableaux, schémas, plans d'architecture...).

Un organigramme informatique est une œuvre de langage artificiel. En effet, la rédaction, comme la composition d'un organigramme répondent à des lois tout à fait explicites, et il est parfaitement possible d'exhiber le dictionnaire des symboles et des mots utilisés dans un organigramme : si cela n'était pas possible alors ce ne serait pas un organigramme informatique².

Examinons maintenant l'objet qui est présenté en page suivante.

1.- JEAN Georges, *Langage de signes, l'écriture et son double*, Découvertes Gallimard, n° 67, Paris, 1989.

2.- Ce pourrait être, par exemple et pour jouer avec les mots, un organigramme d'entreprise.



L'œuvre de langage qui est présentée ici a ceci de commun avec un organigramme informatique qu'elle représente un raisonnement. Cependant si l'on tente sur cette œuvre de langage la procédure de remplacement, elle se déroulera sans aboutir à des primitives.

Cette œuvre de langage est donc exprimée dans un langage naturel.

Dans un organigramme informatique, au contraire, la première étape de la procédure de remplacement fait passer de l'organigramme au logiciel qui a été écrit à partir de cet organigramme puisque, par définition même de l'organigramme informatique, ce dernier est préalable et identique (bien que n'ayant pas la même forme) au logiciel qu'il sert à écrire.

S'il s'agit donc de classer un organigramme, deux hypothèses sont possibles :

- soit l'auteur de l'organigramme ne peut exhiber le logiciel correspondant à l'organigramme ; dans ce cas, l'organigramme fait partie du langage naturel et il est considéré, tout comme notre exemple ci-dessus, comme une forme particulière de description d'un raisonnement.

- soit l'auteur de l'organigramme peut exhiber le logiciel correspondant à cet organigramme ; dans ce cas, l'organigramme fait partie du langage artificiel.

Un problème de classement peut se poser dans le cas où l'auteur d'un organigramme n'exhibe qu'une partie du logiciel correspondant à cet organigramme. Dans ce cas, le juge appréciera, car il est possible que certaines parties d'un

organigramme ne soient pas transformables en logiciel (en l'état actuel des connaissances informatiques), ce qui évidemment retire à cette partie de l'organigramme toute possibilité de protection par la loi de 1985.

Par contre d'autres parties de l'organigramme, même encore non traduites sous forme de logiciel, peuvent l'être techniquement sans problème. Pour ces parties, le juge pourra considérer que l'exhibition du logiciel correspondant n'est pas nécessaire pour amener leur classement en tant que logiciel.

Remarquons cependant que dès qu'apparaît dans un organigramme un sous-ensemble dont la programmation est contestable, le propriétaire de l'organigramme doit impérativement exhiber le logiciel correspondant, faute de quoi l'organigramme, considéré dans sa totalité, perd son statut de logiciel car il ne répond plus à la finalité de faire fonctionner la machine de traitement de données.

Comme nous venons de le voir, certains organigrammes seront donc assimilés à des logiciels alors que d'autres seront simplement considérés comme des expressions en langage naturel.

3.2.3.7.- Les clones logiciels, critères de distinction

Revenons maintenant au problème de la comparaison de deux logiciels ayant les mêmes fonctions.

La créativité de l'analyste (un analyste est un informaticien dont une part importante de l'activité est d'écrire des organigrammes) consiste à concrétiser l'algorithme de la façon la plus concise et la plus efficace possible. De même

qu'étant connu le principe du levier, on peut imaginer une brouette à roue déportée comme en Europe ou une brouette à roue centrale comme en Chine, il y a pratiquement toujours plusieurs façons équivalentes (du point de vue de l'efficacité) de concrétiser un même algorithme.¹

C'est à cause de cette multiplicité des façons de concrétiser un même algorithme que Foxbase n'est pas considéré comme dérivé de DBase et, plus généralement, qu'un clone ne sera pas une dérivation du logiciel original. En effet à partir du moment où les organigrammes sont réellement différents (et la multiplicité des façons possibles d'écrire un organigramme est telle que cela se produira nécessairement si les informaticiens qui ont décidé d'écrire le clone ne sont pas les mêmes que ceux qui ont écrit le logiciel antérieur), le seul élément qui est en amont de l'organigramme, c'est l'algorithme, l'idée non susceptible d'appropriation. Et donc, un élément très important de la comparaison de deux logiciels ayant les mêmes fonctions, la même apparence pour l'utilisateur et écrits dans le même langage de programmation, sera la comparaison des organigrammes que l'on peut trouver dans ce que l'on appelle le dossier d'analyse, qui accompagne nécessairement la rédaction d'un logiciel écrit par des informaticiens professionnels.

Il peut arriver que les dossiers d'analyse soient incomplets, par suite d'une mauvaise maîtrise de l'ingénierie de programmation de l'entreprise. Mais il est totalement exclu que le dossier d'analyse soit incomplet au point que cela rende impossible la comparaison avec un autre dossier d'analyse. On peut considérer aujourd'hui que le refus d'exhibition d'un

1.- Il y a également bien entendu une infinité de façons de la concrétiser de façon inefficace : on peut faire une brouette à un seul bras, sans fond, à roue carrée..

dossier d'analyse est équivalent à un aveu de copie, la copie ayant été suivie d'un maquillage visant à essayer de faire disparaître son aspect servile.

Lorsque les dossiers d'analyse sont disponibles, l'appréciation du juge portera sur deux points :

- la proximité entre les organigrammes. Le juge évaluera ici en quoi les algorithmes de base sont différents et, s'ils sont identiques, en quoi les concrétisations sont différentes.
- si les concrétisations sont identiques, c'est-à-dire dans le cas peu fréquent mais possible où les organigrammes seraient identiques, le juge appréciera les différences nécessairement très minimales dans la traduction en logiciel des organigrammes.

Puisque la comparaison entre deux logiciels fonctionnellement identiques porte principalement sur la composition, on voit que le juge se pose le même problème que lorsque doivent être comparées les compositions d'œuvres en langage naturel. La description d'une composition, d'un texte, d'un dictionnaire, d'un logiciel, peut se faire avec des niveaux de finesse très variables. Le niveau le plus fin c'est l'objet lui-même. Le niveau le plus grossier c'est le nom de l'objet. Pour un livre, la table des matières constitue un niveau intermédiaire à condition toutefois qu'elle soit suffisamment détaillée, c'est-à-dire qu'elle contienne des parties et leurs titres... Dans un dossier d'analyse, les organigrammes sont empilés hiérarchiquement du niveau le plus grossier vers le niveau le plus fin. Chaque niveau de détail constitue une vision partielle mais plus précise que le niveau supérieur.

La comparaison des organigrammes de deux logiciels est donc une tâche lourde (bien que pratiquement réalisable) qui ne doit donc être envisagée que lorsque d'autres moyens n'ont pas fourni la preuve d'une dérivation ou d'une copie.

Le travail considérable que représente un maquillage en profondeur d'un logiciel copié est à peu près équivalent à la réécriture complète du logiciel et demande des compétences comparables sans donner la satisfaction de la réalisation. Aussi, le cas de loin le plus fréquent en matière de copie de logiciel est celui de la copie servile ou aisément détectable (modification des noms des variables, modification des commentaires). Tout le problème alors est de détecter les segments copiés. Il existe aujourd'hui des logiciels destinés à cet objet.

3.2.3.8.- Les logiciels inachevés et les logiciels qui ne fonctionnent pas

Abordons maintenant la question de savoir si un logiciel qui ne fonctionne pas correctement peut cependant être considéré comme relevant néanmoins de la loi de 1985. Le cas est fréquent, pour ne pas dire général.

Il y a tout d'abord le cas des logiciels en cours d'écriture et d'achèvement. Ces logiciels ne sont pas fonctionnels, leur dossier d'analyse peut ne pas être complet. Leur classification en tant que logiciel relève clairement de l'appréciation au cas par cas.

Les juges du Tribunal de commerce de Meaux ont, le 3 novembre 1987, décidé de considérer une œuvre de langage comme un logiciel alors qu'il était en voie d'achèvement ; celui-ci "n'était pas achevé, et, en tout cas pas opérationnel

puisque'il lui manquait sa documentation fonctionnelle" ; il était en conséquence "inutilisable par une autre personne que (son créateur)"¹.

Remarquons que, du point de vue de la procédure de classification qui est le sujet de ce chapitre, un logiciel en cours d'achèvement est, sans contestation possible, une œuvre de langage artificiel et que, si l'on considère au sens strict la définition de la Commission nationale de terminologie de l'informatique, ce n'est pas un logiciel puisque'il ne peut faire fonctionner un ordinateur. Cette position extrême ne peut cependant être tenue à la fois pour des raisons pratiques et pour des raisons théoriques.

Les raisons pratiques sont évidentes : un logiciel perdrait son statut de logiciel à la première défaillance technique, ce qui permettrait à son usager de commencer à en faire des copies multiples. Or, de nombreux logiciels commercialisés provoquent, dans certaines circonstances, des pertes totales de contrôle de la machine.

Les raisons théoriques sont également simples. Il a été démontré par Alan Turing de façon localisée, puis par Kurt Gödel de façon généralisée, qu'il est impossible de démontrer qu'un algorithme permet d'arriver en un nombre fini d'étapes, au résultat qu'on attend de lui. Autrement dit, il n'est pas démontrable qu'un algorithme fonctionne car, si cela était possible, cela voudrait dire que l'on connaît l'algorithme qui permet de vérifier tout algorithme. Or il a été montré que ce problème est indécidable. Puisqu'on n'est jamais sûr qu'un algorithme permet d'atteindre le résultat attendu en un nombre fini d'étapes, cela veut dire qu'il est possible que dans un cas

1.- Lamy droit de l'informatique 1991, n° 1309.

particulier du problème que l'algorithme est censé résoudre, il faudra un nombre infini d'étapes, ce qui autrement dit signifie qu'on n'arrive jamais au résultat. En termes informatiques, un ordinateur qui serait dans ce cas particulier, commencerait un calcul de longueur infinie. L'exemple typique d'un tel calcul est la boucle sans test de sortie. Lorsqu'il s'agit d'écrire un logiciel un peu compliqué, c'est-à-dire qui dépasse une dizaine de lignes de programmes, l'algorithme sous-jacent commence à être un peu compliqué également. Comme le programmeur ne dispose pas d'un logiciel qui lui garantirait la validité de son algorithme, il en est réduit à explorer lui-même toutes les configurations possibles afin de vérifier que son algorithme, et donc son logiciel, fonctionne bien. Malheureusement, il oublie un nombre d'autant plus grand de configurations que le logiciel est compliqué. Ces configurations ne se produisent parfois que très longtemps après que le logiciel ait été commercialisé. Leur correction est la raison principale (outre les améliorations fonctionnelles) de l'existence de versions successives d'un même logiciel. Et en tout cas, il reste définitivement indémontrable qu'un logiciel fonctionne totalement correctement. On peut seulement dire que le nombre de situations de bon fonctionnement du logiciel augmente régulièrement au fur et à mesure que les erreurs sont détectées et corrigées. On peut dire également que lorsqu'un logiciel a été diffusé à plusieurs centaines de milliers d'exemplaires, qu'il est utilisé par plusieurs millions d'utilisateurs sur des machines variées, alors la probabilité qu'il contienne une erreur qui empêchera réellement un nouvel utilisateur de s'en servir (et un ordinateur de fonctionner) est très faible. Elle peut même être nulle, car il est tout à fait possible, même si cela n'est pas démontrable, que l'algorithme fonctionne entièrement correctement, dans toutes les situations possibles, en un nombre fini d'étapes.

Il est donc indémontrable théoriquement comme pratiquement qu'un logiciel répond strictement et entièrement à toutes les définitions, françaises comme étrangères, qui contiennent un critère de finalité.

Selon l'article 7 de la loi du 11 Mars 1957, "l'œuvre est réputée, créée, indépendamment de toute divulgation publique, du seul fait de la réalisation, même inachevée, de la conception de l'auteur". Cette disposition, dans le cas des logiciels, rejoint les considérations pratiques et théoriques que nous venons d'évoquer.

Ainsi la question n'est pas de savoir si une création inachevée peut-être protégée mais de savoir à partir de quel niveau d'achèvement une création peut devenir œuvre logicielle. Pour toutes les raisons qui viennent d'être évoquées, on peut dire que l'œuvre est assimilée à un logiciel aux différents stades de la création. Notre position est, sur ce point, en plein accord avec celle qui est exprimée par les auteurs du Lamy-Droit de l'informatique. Ainsi "le logiciel doit être reconnu comme tel au fur et à mesure de son écriture, aussi imparfait soit-il et quelles que puissent être les nécessités de revenir sur un premier jet insatisfaisant. Cela peut répondre (...) à l'esprit de la loi de 1957 considérée globalement et évite, qui plus est, une césure aussi mal venue que non maîtrisable entre un stade qui serait "pré-logiciel" (...) et un stade ultérieur qui pourrait seul être tenu pour logiciel"¹.

Mais, tout comme il semble difficile d'accepter que les deux premiers mots d'un roman constituent le roman inachevé, on a du mal à accepter qu'une seule ligne de programme soit un logiciel de gestion de central téléphonique inachevé. Un expert

1.- Lamy droit de l'informatique 1991, n° 1309.

en informatique dira en effet tout de suite qu'un logiciel de gestion de central téléphonique peut difficilement comprendre moins de 100 000 lignes de programme. Rappelons à cette occasion que, bien entendu, si la ligne de programme exhibée est accompagnée de l'organigramme complet et techniquement valide permettant d'écrire facilement les lignes de programme manquantes, alors, au contraire, il semble peu contestable d'affirmer que "l'œuvre logicielle" existe même si elle est inachevée. Le juge appréciera le degré d'achèvement d'une création de langage artificiel qui lui est proposée afin de déterminer si, éventuellement, elle peut être considérée comme une œuvre logicielle.

3.2.3.9.- La documentation

Une autre question est souvent posée par les juristes à propos des logiciels : savoir si la documentation associée au logiciel est un élément du logiciel.

Plusieurs définitions du terme "logiciel" incorporent la documentation. Ainsi la Commission nationale de terminologie de l'informatique définit le logiciel comme étant "l'ensemble des programmes, procédés et règles et éventuellement de la documentation...". Ce que nous avons dit précédemment éclaire ce problème d'un jour qui, sans être nouveau, permet néanmoins de donner quelques bases supplémentaires pour des décisions de justice.

Faisons au préalable deux petites remarques :

Remarque n°1 : Une œuvre, y compris de langage, peut tout à fait être présentée par segments, chaque segment ayant un support différent. Ainsi il est fréquent en littérature, qu'un

roman soit présenté en plusieurs tomes. L'œuvre sera considérée comme unique même si le tome 1 est édité dans la Pléiade et le tome 2 en Livre de poche. Strictement rien ne s'oppose donc à ce qu'un logiciel soit exhibé pour partie sur disquette, pour partie sur listing et, pourquoi pas, sur un support ayant l'apparence d'un livre. Remarquons d'ailleurs que, de plus en plus fréquemment, la documentation destinée à l'utilisateur est non seulement fournie sous forme d'un livre, mais aussi intégrée dans le corps même du logiciel, où elle peut être activée pour apparaître à l'écran par l'une des commandes du langage de commande. La documentation fait à ce titre partie du logiciel comme n'importe quel autre message d'alarme ou de dialogue avec l'utilisateur. Elle est, de plus, fournie sur le même support que le reste du logiciel.

Remarque n° 2 : L'unicité d'une œuvre de langage n'est pas remise en cause par le fait que sa forme ne soit pas homogène. Ainsi de nombreux ouvrages de Jacques Prévert sont pour parties constituées de poèmes et pour partie constituées de récits ou de nouvelles. De même l'ouvrage d'Italo Calvino, *Le château des destins croisés*¹, est originellement constitué d'un texte et de dessins représentant des configurations de tarot. Notre procédure de remplacement appliquée à cet ouvrage ne fonctionne pas.

En effet, pour toute la partie textuelle elle montrera bien qu'il s'agit de langage naturel, par contre, elle ne pourra être appliquée au dessin des configurations de tarot puisque ces objets ne sont pas des œuvres de langage. Il ne viendrait cependant à l'idée de personne de dire que ce livre n'est pas un livre sous prétexte qu'il contient des illustrations.

1.- CALVINO Italo, *Le château des destins croisés*, Editions du Seuil, Collections points, 1976.

Anecdotiquement remarquons que nous pourrions exhiber des tableaux de peintres célèbres contenant des textes et dont personne ne contesterait que ce sont des tableaux.

A mi-chemin des livres contenant des illustrations et des tableaux contenant des textes, il y a les bandes dessinées et aussi la célèbre œuvre de René Magritte intitulée "ceci n'est pas une pipe", à propos de laquelle le philosophe Michel Foucault disait : "rien de tout cela n'est une pipe ; mais un texte qui simule un texte ; un dessin d'une pipe qui simule un dessin d'une pipe ; une pipe (dessinée comme n'étant pas un dessin) qui est le simulacre d'une pipe (dessinée à la manière d'une pipe qui ne serait pas elle-même un dessin). Sept discours dans un seul énoncé"¹.

Nous avons démontré précédemment que la présence d'une fable de Jean De La Fontaine dans un logiciel ne brisait pas son unité. La documentation d'un logiciel, tout comme les commentaires dans le corps même du texte du logiciel, n'est absolument pas utilisée par la machine lors de l'exécution du programme au sens strict des définitions d'un logiciel. Les commentaires, tout comme la documentation associée, sont donc superflus. Nous avons cependant démontré l'ineptie d'une procédure de décomposition lorsqu'il s'agit de classer une œuvre de langage. Cette considération réintègre donc sans discussion possible les commentaires ainsi que la documentation, lorsqu'elle fait partie du corps même du logiciel.

Si la documentation associée au logiciel ne fait pas partie du corps même du logiciel, mais est fournie dans des documents

1.- Texte cité par Jean-Luc RISPAIL - In : *Les surréalistes, une génération entre le rêve et l'action*, Gallimard, Paris, 1991.

séparés (livre, microfiches...), la procédure de remplacement des mots par leur définition ne donnera pas le même résultat pour le corps même du logiciel et pour la documentation séparée, puisque dès la première phrase de l'ouvrage qui constitue le support de la documentation, on sera dans un cas de langage naturel.

Notre position est de considérer que la documentation séparée du logiciel doit être jugée partie intégrante du logiciel dans le cas où le reste fait explicitement partie du cadre de l'affaire en cours de jugement. En effet, si la documentation est superflue pour l'exécution du langage artificiel par la machine, par contre elle est absolument indispensable pour la réalisation du traitement de données. Sans elle, en effet, un humain non formé ne pourra déclencher l'exécution par la machine.

Nous pensons qu'il convient de distinguer la documentation qui est effectivement indispensable à un humain pour réaliser le traitement de données, de la documentation qui ne l'est pas, par exemple les exercices pédagogiques et autres manuels secondaires.

La documentation, non indispensable (ainsi que la documentation indispensable, dans le cas où l'affaire en cours de jugement porte uniquement sur la documentation, sans porter sur le corps du logiciel lui-même), relèvera exclusivement de la loi de 1957. Notre position concernant la documentation indispensable dans le cas où un litige ne porterait que sur elle, sans porter sur le corps même du logiciel, s'explique très simplement par le fait que notre procédure de remplacement désigne cet objet comme une œuvre de langage naturel, sans qu'aucune pondération résultant de la présence du corps du logiciel vienne modifier le

classement. Nous avons déjà traité ce cas plus haut en évoquant le manuel de programmation qui contiendrait des exemples de logiciel.

3.2.4.- Systèmes experts et bases de connaissances

Pour commencer, précisons les objets sur lesquels nous travaillons. Cette précision est indispensable car les termes de "système expert" et de "base de connaissances" recouvrent suivant les auteurs des réalités fort différentes.

Remarquons qu'une fois de plus il n'est pas question ici de rentrer dans des considérations techniques pointues, mais de s'efforcer d'utiliser la langue avec rigueur.

3.2.4.1.- Proposition de définition d'un système expert

Par "système expert", on désigne habituellement un objet avec lequel l'utilisateur engage des interactions (des actes de communication écrits, parlés, visuels mais pourquoi pas également graphiques, tactiles ou olfactifs) qui visent à transmettre des connaissances du système expert vers l'utilisateur ou à faire émettre par le système expert des avis qui seraient comparables à ceux qu'émettrait, dans les mêmes circonstances, un expert humain.

Cette définition est d'une généralité extrême et, par là même, elle est très imprécise. Une maman vis-à-vis de son bébé répond à cette définition du système expert. Donnons donc quelques critères distinctifs supplémentaires.

Tout d'abord un système expert n'est évidemment pas un être vivant (c'est-à-dire humain, animal ou végétal), mais un système expert n'est pas non plus une machine. Il est

absolument indispensable en effet de distinguer, d'une part, l'objet constitué d'une boîte noire que l'on branche sur le courant pour obtenir des connaissances et, d'autre part, les œuvres de langage éventuellement contenues dans cette boîte noire. L'exemple le plus simple de cette indispensable distinction est le four à micro-ondes, boîte noire dont la fonction est de cuire et de réchauffer des aliments, qui ne doit pas être confondu avec le logiciel contenu dans la mémoire de l'ordinateur qui constitue une des pièces du four à micro-ondes.

Revenons au système expert. Il existe déjà aujourd'hui et il va exister en nombre croissant des objets, généralement portables, qui auront pour fonction d'assister l'humain dans des activités cognitives : secouriste électronique à placer dans les vide-poches des voitures, à côté du système de cartographie, du traducteur de poche ou du détecteur d'odeur de fraises des bois (très utile lors des promenades en forêt)... Ces objets existent déjà et la seule chose qui s'oppose à leur transportabilité est, pour très peu de temps encore, leur dimension trop importante. Il n'y a, pour deux motifs principaux, aucune raison de qualifier ces objets de logiciels.

- La première de ces raisons est qu'un logiciel a été défini comme étant une œuvre de langage constituée de caractères ou d'idéogrammes. Or ces objets ne sont pas des œuvres de langage mais des assemblages de métal et de plastiques, des machines.

- La deuxième raison est que strictement rien ne permet extérieurement de dire de quelle façon est réalisée la fonction de traduction, cartographie, assistance, médicale... par ces machines. Supposer que la fonction est réalisée par ordinateur est purement gratuit, même si c'est temporairement exact à la

fin du vingtième siècle. On a ainsi vu, à la fin du dix-huitième siècle, une machine à jouer aux échecs (archétype du système expert) dans laquelle se cachait un nain. On sait aujourd'hui fabriquer des détecteurs d'odeur de fraise des bois qui ne contiennent pas le moindre élément informatique (ils fonctionnent par piégeage moléculaire). Il faut donc s'interdire absolument d'identifier le concept de machine manipulant de la connaissance avec le concept de logiciel.

Dans ce qui suit, nous ne désignerons par le terme "système expert" que des objets qui sont des œuvres de langage, quel que soit, bien entendu, le support sur lequel ces œuvres sont inscrites.

On distinguera donc le four à micro-ondes du texte du logiciel inscrit dans ses mémoires et le boîtier du jeu d'échecs des textes (logiciel, description de parties...) contenus dans ses mémoires.

Sans chercher, pour l'instant, à fournir une définition précise, nous considérerons donc ici qu'un système expert est constitué de l'ensemble des textes contenus dans les mémoires d'une machine qui a pour fonction d'assister son usager dans des activités de raisonnement.

3.2.4.2.- Les systèmes experts sont-ils des logiciels ?

Un débat a fortement animé la communauté des juristes ces dernières années : il s'agissait de savoir si un système expert était ou n'était pas un logiciel. Dans la réponse que nous proposons à cette question, nous ne reviendrons pas ici en détail sur l'architecture des systèmes experts car ce serait commettre une erreur identique aux ouvrages juridiques dont

les explications techniques se caractérisent autant par leur incomplétude que par leur absence de clarté.

Nous ne raisonnerons ici que sur les trois éléments qui sont habituellement utilisés pour caractériser des systèmes experts :

- la séparation entre moteur d'inférence et base de connaissances,
- la forme déclarative des bases de connaissances,
- un raisonnement automatique sur des données incomplètes.

Nous sommes absolument conscient du fait que ces trois éléments ne sont pas caractéristiques des systèmes experts.

En effet :

- on trouve de nombreux logiciels dans lesquels il y a séparation entre les programmes et les données.
- on trouve également de nombreux systèmes d'information dans lesquels les données sont sous forme déclarative.
- on trouve enfin de nombreux logiciels qui fonctionnent (même s'ils ne donnent pas un bon résultat) y compris lorsque les données à traiter sont incomplètes.

Cette imprécision résulte du fait que la notion même de système expert est imprécise.

Pour commencer, notons qu'il est clair qu'un système expert, pris dans son ensemble (moteur d'inférence + base de connaissances + base de faits) et défini ainsi que nous venons de le faire comme une œuvre de langage, est un logiciel. L'argumentation est la même que celle qui a été présentée dans notre exemple du logiciel qui imprimait une fable de Jean De La Fontaine. En effet, quels que soient les éléments qui la

composent, cette œuvre de langage a donc bien pour finalité de faire fonctionner un ordinateur et est bien interprétable par la machine.

Nous avons déjà vu, à partir de plusieurs exemples, qu'il était stupide de décomposer cette œuvre en sous-éléments pour tenter de qualifier l'œuvre dans son ensemble. Cela étant, on peut se poser la question de savoir si chacun des sous-ensembles constituant un système expert conserve un statut de logiciel lorsqu'il est séparé de l'œuvre complète.

Nous allons, évidemment, maintenant considérer d'une part la base de connaissances et la base de faits et, d'autre part, le moteur d'inférence et l'environnement de développement. Dit d'une autre façon, nous examinerons séparément les bases de faits et les bases de connaissances qui sont rédigées par l'utilisateur après l'achat du moteur d'inférence et de l'environnement de développement. Il n'y a, dans la littérature juridique, aucune contestation sur le fait que le système expert "vierge" est un logiciel. C'en est un exactement au même titre qu'un tableur, un S.G.B.D. ou un logiciel de traitement de texte. La différence entre un système expert vierge et l'un de ces trois logiciels réside simplement dans le fait que le système expert vierge est destiné à manipuler une classe d'œuvres qui ne sont ni des tableaux, ni des fichiers, ni des textes quelconques.

Examinons maintenant la base de connaissances et la base de faits. Leur rédaction doit très strictement respecter les contraintes syntaxiques. L'exemple le plus souvent donné de telles contraintes est la forme SI...ALORS...SINON... que doivent respecter les énoncés de certaines familles de règles dans les bases de connaissances. Il y a bien d'autres contraintes qui sont particulières à chaque système expert "vierge". Comme le

fait remarquer Jean Deschamp, cité par Jean-Michel Piot¹, "à chaque système expert vierge, correspond une syntaxe particulière des bases de connaissances qui seront rédigées avec ce système expert vierge". Il se trouve que, tout comme dans le thésaurus, les syntaxes imposées par les concepteurs de systèmes experts vierges empruntent largement leurs formes aux graphies du langage naturel des humains. Ceci a pour conséquence qu'un humain, non informaticien et non linguiste, peut avoir l'impression que le texte d'une base de connaissances est du langage naturel, tout comme la lecture de certains logiciels écrits dans des langages de programmation très évolués peut donner l'impression qu'ils sont écrits en langage naturel.

Nous avons déjà discuté de l'impossibilité de déterminer la qualité d'un texte sur le seul critère d'apparence. Le texte d'une base de connaissances ou d'une base de faits n'est absolument pas une expression du langage naturel. C'est une expression du langage artificiel, comme le prouve l'application de notre procédure de remplacement.

Par contre, la base de connaissances n'est pas un logiciel. En effet, bien que rédigée dans un langage artificiel interprétable par un ordinateur, le texte de la base de connaissances ne décrit pas une séquence d'opérations que la machine devra exécuter. En théorie, l'ordre des phrases dans une base de connaissances est sans importance. Le texte de la base de connaissances ne décrit pas un algorithme. Considérée séparément du système expert qui a servi à la développer, une base de connaissances est une œuvre de langage artificiel qui relève de la loi du 11 mars 1957. Tout comme un thésaurus,

1.- PIOT Jean-Michel, *Les créations informatiques complexes* - Mémoire de D.E.A., Montpellier, 1987.

cette œuvre de langage artificielle n'est, en effet, pas exécutable par une machine. Elle constitue une partie des données utilisées lors de l'exécution par la machine du moteur d'inférence qui, pour sa part, est un logiciel, tout comme est un logiciel la requête de l'usager au système expert.

Nous avons vu que la syntaxe d'une base de connaissances est totalement liée au système expert vierge qui a servi à la développer. Il n'existe pas actuellement de langage artificiel général permettant de décrire des connaissances de façon telle que la description soit utilisable par des systèmes experts différents. Cela peut cependant arriver un jour. Dans le cas d'une copie de la base de connaissances (sans copie du système expert vierge qui a servi de support à sa conception), le copieur n'aurait pas lieu d'être poursuivi tant qu'il ne fait pas d'usage public de sa copie. En effet, la base de connaissances seule, expression dans un langage artificiel mais n'ayant pas le statut de logiciel, relève à notre avis de la loi du 11 mars 1957, qui autorise des copies pour usage privé du copiste. Notre position à propos de cet objet est identique à celle qu'il nous semble falloir adopter vis-à-vis des textes contenus dans les bases de données : ces textes¹ ne sont pas des logiciels car ils ne servent pas à faire fonctionner des machines de traitement de données.

3.2.5.- Conclusions partielles

Tout au long de ce chapitre, lorsque nous avons décomposé un logiciel en sous-éléments, pour juger non pas le logiciel en tant que tel, mais chacun des éléments particuliers, nous

1.- Sauf cas très particulier des bases de données qui contiennent les listages de logiciels.

avons pu obtenir des jugements classificatoires fort différents :

1- dans l'exemple de la fable de La Fontaine, nous avons trouvé un sous-ensemble qui est une œuvre de langage naturel protégé par la loi de 1957 ;

2- dans le cas des commentaires et des remarques intégrés dans le corps des programmes, nous trouvons des textes en langage naturel qui ne peuvent pas être séparés véritablement du logiciel et qui relèvent tout comme lui de la loi de 1985 ;

3- nous avons ensuite rencontré le cas des documentations indispensables associés au logiciel. Ces textes en langage naturel peuvent avoir tantôt un statut de logiciel, tantôt un statut d'œuvre non logicielle ;

4- tout comme la documentation, les bases de connaissances, œuvres de langage artificiel, peuvent avoir un statut de logiciel lorsque le litige porte sur l'ensemble du système expert et un statut d'œuvre non logicielle lorsqu'elles sont considérées isolément.

La variété des statuts des sous-ensembles qui sont susceptibles de composer un logiciel constitue, à notre avis, une démonstration solide de l'erreur, fréquemment commise, qui consiste à utiliser la méthode de la décomposition pour déterminer si un objet est un logiciel. Ce résultat, à notre connaissance, n'a jamais été exposé dans la littérature juridique.

Par ailleurs, nous sommes bien conscients que la procédure de classification que nous proposons à partir de la méthode des

remplacements n'a pas la solidité d'une démonstration qui serait faite à l'intérieur d'un cadre formel solide. Comme nous l'avons vu au début de notre démonstration, celle-ci repose sur une différenciation entre les langages qui possèdent une sémantique et les langages qui n'en possèdent pas. Or la question de l'existence même du sens (et donc de la sémantique) fait l'objet d'un débat non achevé à ce jour. Ce débat dépassant le cadre de notre travail, nous nous sommes placé dans le cadre le plus largement accepté. Dans ce cadre, l'application aussi rigoureuse que possible de notre méthode nous a permis de proposer, pour nombre de cas, des classifications qui, d'une part, rencontrent les intuitions des meilleurs juristes et, d'autre part, s'inscrivent toutes dans une même cohérence. Dans de nombreux autres cas, il nous semble que l'application de la méthode nous a permis de lever des indéterminations qui subsistaient encore.

Bien entendu, des avancées scientifiques et techniques, qui remettraient notre méthode ou nos conclusions classificatoires en cause, sont possibles et nous les attendons avec impatience. En effet, la possibilité de voir un jour des machines manipuler la langue naturelle avec une dextérité humaine nous intéresse au plus haut point. En attendant ces événements, poursuivons maintenant notre activité classificatoire en appliquant les résultats précédents aux produits issus des industries de la langue.

3.3.- Une classification des produits issus des industries de la langue

Parmi les très nombreux produits issus des industries de la langue, nous allons maintenant rechercher ceux auxquels on peut attribuer un statut de logiciel. Les chapitres précédents sont suffisamment détaillés pour que cette recherche puisse être menée à bien rapidement. Aussi nous ne reprendrons pas dans ce qui suit le détail des argumentations et nous renverrons au chapitre précédent où elles sont présentées.

Reprenons tout d'abord la distinction entre machines et œuvres de langage. Le *Speak and Spell* vendu par la société Texas Instrument¹ est une machine. Tout comme le traducteur de poche, cette machine contient dans ses mémoires des œuvres de langage qu'éventuellement nous qualifierons de logicielles. Mais, comme nous l'avons vu, il ne faut pas confondre la machine et les œuvres de langage qui sont dans la machine et qui servent, éventuellement, à la faire fonctionner. Dans ce qui suit, nous ne chercherons donc à qualifier que des œuvres de langage.

3.3.1.- Les dictionnaires informatisés

Nous ne parlons ici que des données lexicographiques ou terminologiques à l'exclusion des logiciels qui servent à les créer ou à les manipuler.

1.- Le *Speak and Spell* est un jouet connu également sous le nom de *Dictée magique*. Il permet à un enfant d'apprendre l'orthographe. Il contient un module de synthèse de la parole et un petit dictionnaire qui dans les mémoires de la machine associe à chaque mot d'une part son orthographe et d'autre part sa décomposition phonétique.

Les bases de données lexicographiques sont des œuvres de langage naturel, bien qu'ayant un support qui n'est pas toujours du papier ; notre procédure de remplacement fait clairement apparaître que ces textes, tout comme les textes de dictionnaires destinés à des humains qui sont sur des supports papier, ne sont pas transformables en objets composés de primitives d'un langage artificiel.

3.3.2.- Les dictionnaires électroniques

Les dictionnaires électroniques sont destinés à être utilisés par des machines. La forme de ces textes est strictement contrainte. Le plus souvent d'ailleurs ce ne sont pas des textes au sens classique du terme mais des tableaux de chiffres, des matrices de valeurs binaires et des listes constituées de codages. Les règles d'écriture de ces dictionnaires électroniques n'ont absolument rien à voir avec des règles d'écriture du langage naturel. Des procédures de remplacement appliquées à ces dictionnaires électroniques ramènent des primitives qui ne sont pas des instructions à des ordinateurs.

Les dictionnaires électroniques sont donc des œuvres de langage artificiel sans être des logiciels. Bien entendu cette proposition de classification ne vaut que pour l'objet considéré isolément. En effet, un dictionnaire électronique intégré dans un logiciel de vérification orthographique a tout comme l'objet dans lequel il est inclus, le statut de logiciel.

3.3.3.- Les corpus de références

Cet ensemble organisé d'extraits de textes, de discours, de dialogues, d'enregistrements de sons... se place évidemment dans la catégorie des œuvres rédigées en langage naturel.

3.3.4.- Les œuvres obtenues à la sortie d'un traitement par une machine

Ces œuvres auront, dans les cas les plus fréquents, un statut d'œuvre de langage naturel.

Ce sera le cas par exemple du texte issu d'un lecteur optique ou d'une machine à écrire à entrée vocale. Ce sera aussi le cas des textes issus d'un système de correction orthographique ou d'un système de traduction assisté par ordinateur.

Dans certaines applications, toutefois, il pourra s'agir de textes rédigés en langage artificiel : une liste de mots-clés extraits d'un thésaurus, les messages de la météorologie canadienne.

Sauf cas très exceptionnel (le texte d'un logiciel saisi dans un système de traitement de texte), les œuvres issues du traitement automatique dans le domaine des industries de la langue, n'auront pas le statut de logiciel.

3.3.5.- Les logiciels

Dans la plupart des machines, il y aura des œuvres de langage qui seront des logiciels.

On peut ainsi acheter des logiciels (achat des objets physiques)¹ de vérification orthographique (comme HUGO, GRAM-R, Sans Faute, Orthogiciel...), des logiciels de traduction

1.- Nous rappelons que "la propriété incorporelle (...) est indépendante de la propriété de l'objet matériel" (article 29, alinéa 1 de la loi du 11 mars 1957).

assistée par ordinateur (Weidner, Alps, Systran, Titus...), des logiciels d'indexation automatique et de gestion de documentation (Darwin, Spirit...). Ces objets sont vendus sur support magnétique ou optique et sont accompagnés de leur documentation indispensable. Lorsqu'on examine ces objets en essayant de les décomposer en sous-ensembles, on constate souvent qu'ils contiennent :

- des sous-ensembles rédigés en langage artificiel et qui se révèlent être des logiciels parce que la procédure de remplacement ramène finalement des instructions et des commandes destinés à un ordinateur,
- des sous-ensembles rédigés en langage artificiel et qui ne sont pas des logiciels mais des dictionnaires électroniques, des thésaurii...
- des sous-ensembles rédigés en langage naturel, comme leur documentation en ligne, ou bien des lexiques et dictionnaires destinés à des humains, des encyclopédies, etc.

Le fait que certains des sous-ensembles composant un logiciel ne soient pas, lorsqu'ils sont considérés isolément, des logiciels ne change rien au fait que l'objet dans son ensemble est un logiciel. Dans la plupart des cas, cet objet est d'ailleurs commercialisé en tant que logiciel. C'est le cas de tous les logiciels qui viennent d'être cités.

Parfois cependant, le vendeur préfère, pour des raisons commerciales, ne pas donner à cet objet un statut de logiciel mais un statut d'ouvrage relevant uniquement du droit commun de la propriété littéraire et artistique (loi de 1957). Ainsi un éditeur de dictionnaires (papier) qui commercialise une disquette contenant un dictionnaire de biologie (destiné à être

utilisé par des humains) et, sur cette même disquette, le logiciel permettant d'accéder aux données, présentera l'objet comme un dictionnaire et non comme un logiciel. De même, l'encyclopédie Grolier sur CD-ROM est présentée comme une encyclopédie alors que notre méthode d'analyse concluerait que c'est un logiciel contenant des données encyclopédiques. Notre méthode, strictement appliquée ne classerait cette œuvre comme encyclopédie que s'il n'y avait sur le support optique que les données, mais ce n'est pas le cas : sur les CD-ROM de Grolier, il y a un logiciel très complexe dont l'exécution par une machine permet d'accéder aux données.

Sur certaines pochettes de CD-ROM, il est fait mention d'un double copyright : l'un pour les données, l'autre pour le logiciel de pilotage du CD-ROM. C'est le cas pour le CD-ROM "Politique et Société" : les données appartiennent à la Documentation Française et le logiciel à la Société Act Informatique. On peut émettre l'hypothèse que, pour les producteurs de ce CD-ROM, cette double mention a pour objet de se préserver dans le plus grand nombre possible de situations.

Bibliographie

Bibliographie

AMIEL Philippe, *Quand le dictionnaire s'automatise*, in : "Les intelligences artificielles" - revue Sciences et Avenir, n° 53 spécial, pages 94-100, 1985.

AMIEL Philippe. *Vers les dictionnaires électroniques*. In Journées européennes de la traduction professionnelle, Paris 25-26 mars 1987, Encrages, Université Paris 8, n°17, 1987.

AUGER Pierre, *La notion d'industrie de la langue, essai de typologie*, In : les industries de la langue : un domaine à la recherche de lui-même, Conseil de la langue française, Québec, 1990.

BALPE Jean-Pierre, *Initiation à la génération de textes en langue naturelle*. Editions Eyrolles, Paris 1987.

BAR HILLEL Yeoshua, *indexical expressions, aspects of langage*, Jerusalem, The Magnes Press, The Hebrew University, 1970.

BERTIN Gérard et LAMBERTERIE (de) Isabelle, *La protection du logiciel, enjeux juridiques et économiques*, LGDJ, Paris 1985.

BERTRAND André, *Protections juridiques du logiciel*, Editions des Parcs, 1986.

BERTRAND André, *La protection juridique des jeux vidéo (de l'enjeu économique et social à la problématique juridique)*, Expertises n°56, Paris, 1983.

BORILLO Mario, *L'informatique pour les sciences de l'homme*, Editions Pierre Mardaga, Bruxelles, 1982.

BOULAYE Guy, *Eléments de droit pour informaticiens*, Editions Ellipses, Paris 1990.

BRIAT Martine, *Droit de l'informatique*, 1986, n°2.

CALVINO Italo, *Le château des destins croisés*, Editions du Seuil, Collections points, 1976.

CARRE René, DEGREMONT Jean-François, GROSS Maurice, PIERREL Jean-Marie, SABAH Gérard, *Langage humain et machine*, Presses du CNRS, Paris, 1991.

CASSEN Bernard, *Les industries de la langue, un grand enjeu culturel, scientifique et technologique pour la France*, Rapport présenté à M. Hubert Curien, Ministre de la Recherche et de la Technologie - Paris, 7 Novembre 1985.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

CASSEN Bernard, DEGREMONT Jean-François et alt., *Les industries de la langue, enjeux pour l'Europe*, Actes du Coloque de Tours, numéro spécial de la revue *Encrages*, N° 16, France, Paris, Université Paris 8.

CASSEN Bernard (dir.) et alt. *Quelles langues pour la science ?* - Paris, Editions de la Découverte, 1990.

CALLIOPE (aut. col.), *La parole et son traitement automatique*, Jean Pierre Tubach (éd.), Collection Technique et scientifique des Télécommunications, Masson, Paris, 1989.

CATTENAT Annette, PAUL Gérard, *Etude de quelques langages documentaires*, Université Paris VIII, 1977.

CHAUDIRON Stéphane, CONFLAND Daniel, *Le génie linguistique : contexte et développement*, Actes du Colloque "Génie linguistique 91", Versailles, France, 16 et 17 Janvier 1991.

CHAUMIER Jacques, *Les techniques documentaires*, PUF, Paris, que sais-je ?, n°1419.

CHERPILLOD Yvan, *L'objet du droit d'auteur*, Cedidac (Centre du droit de l'entreprise de l'Université de Lausanne), diffusé en France par LITEC, Lausanne, 1985.

CHIVA Nathalie, mémoire de DEA, *Un problème d'inversion de textes*, publié par le CERIL (Centre d'études et de recherches en informatique linguistique), Université Paris 7, septembre 1986.

CHOISY Maris-Gaëlle, *Banques de données, aspects contractuels*, Agence de l'Informatique, Paris, 1983.

CLERMONT Pierre *le pari perdu de la traduction automatique* - Société Magazine, n°4, Juin 1990, page 27.

COLOMBET Claude, *Propriété littéraire et artistique et droits voisins*, Editions Dalloz, Paris, 1988.

CONFLAND Daniel, CHAUDIRON Stéphane, *Le génie linguistique : contexte et développement*, Actes du Colloque "Génie linguistique 91", Versailles, France, 16 et 17 Janvier 1991.

DANLOS Laurence, *La génération automatique de textes*, Editions Masson, Paris, 1985.

DE BESSE Bruno, *Termes techniques nouveaux*, Feutry Editeur, Paris, 1982.

DEGREMONT Jean-François, *L'émergence d'une industrie de la langue*, In : Brises, Paris, Editions du CDSH, n°4, Avril 1984.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

DEGREMONT Jean-François, *Ethnométhodologie et innovation technologique : le cas du traitement automatique des langues naturelles* - Doctorat d'Ethnologie-Anthropologie, Université Paris VII, 1989.

DENIS Sabine, POULLET Yves et TUNIS Xavier, *Banques de données, quelle protection juridique ?*, Cahiers du CRID, Editions Story Scientia, Namur, 1988.

DESBOIS Henri, *Le droit d'auteur en France*, Paris, Editions Dalloz, 1978, n°1.

DREYFUS Hubert, *Intelligence artificielle, mythes et limites*, Paris, Editions Flammarion, 1984, 443 pages.

DUBOIS Jean et DUBOIS-CHARLIER Françoise, *Incomparabilité des dictionnaires*, in : Dictionnaires électroniques du français, sous la direction de Blandine Courtois et Max Silberstein, Langue Française numéro 87, Editions Larousse, Paris, septembre 1990.

DUCROT Oswald et TODOROV Tzvetan, *Dictionnaire encyclopédique des sciences du langage*, collection Point, Editions du Seuil.

FANTAPIE Alain et de BRULE Marcel (sous la direction de), *Dictionnaire des néologismes officiels*, Franterm, Paris, diffusé par Nathan, 1984.

FERRETTI Marc et CINARE François, *Synthèse et reconnaissance de la parole*, Editest, Paris, 1983.

FRANÇON André, *Revue trimestrielle de droit commercial*, 1982, 431.

FRANÇON André, *La propriété littéraire, artistique et industrielle*, Les cours du droit, LGDJ, Paris, 1989.

GASNAULT Jean et VIGNAL Jean-Christophe, *Une protection juridique des systèmes experts*, Mémoire de DEA, Université de Montpellier, 1986.

GAUDIN Thierry, *Les métamorphoses du futur*, collection CPE-Economica, Editions Economica, Paris, France, 1988.

GARFINKEL Harold, *Studies in ethnomethodology*, Prentice Hall Inc., New-York, U.S.A., 1967.

GAUDRAT Philippe, *Droits et obligations de l'utilisateur d'une banque de données*, In Appropriation et circulation de l'information, Brises, CNRS, novembre 1988, n°12.

GAUDRAT Philippe, *Le statut juridique des idées*, in Appropriation et circulation de l'information, Brises, CNRS, Novembre 1988, n°12.

GILLES Bertrand, *Histoire des techniques*, NRF-La Pléiade, Paris, 1978.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

GINSBURG Jane, *Banque de données et droit d'auteur aux Etats-Unis* - In : Banques de données et droit d'auteur - Colloque IRPI, Paris, 27 novembre 1986, LITEC, 1987.

GUENARD Françoise et LELIEVRE Gérard, *Penser les mathématiques* - Paris, Editions du Seuil, 1982.

GODET Michel, *Prospective et planification stratégique*, CPE-Economica, Editions Economica, Paris, 1985.

GOUNOT Jean-François, *La protection des jeux vidéo et des ludiciels*, Mémoire de DEA, Université de Montpellier, 1986.

GROSS Maurice et LENTIN André, *Notions sur les grammaires formelles*, 2ème Edition, Gauthier/Villars - Paris 1970.

GROSS Maurice, *La construction de dictionnaires électroniques*, in Annales des Télécommunications, tome 44, n°1-2, France, janvier-février 1989.

HATON Jean-Pierre, CAELEN J., GAUVIN J.L., PERENNOU G., PIERREL J.M., *La reconnaissance automatique de la parole*, Dunod, Paris, 1991.

HUET Jérôme, *La protection des logiciels, pratiques passées, régime actuel et difficultés futures*, in Expertises, n° 86, Août 1986.

HUET Jérôme, *La modification du droit sous l'influence de l'informatique : aspects de droit privé*, JCP, 1983-1, 3095.

HUET Jérôme et LUCAS André, *La protection des logiciels*, in "Droit d'auteur et droits voisins, la Loi du 3 juillet 1985", Colloque de l'IRPI, novembre 1985, Edition LITEC, Paris, 1986.

JEAN Georges, *Langage de signes, l'écriture et son double*, Découvertes Gallimard, n° 67, Paris, 1989.

JONQUERES Jean, (Rapport du Conseiller), *Revue de droit de la propriété industrielle*, 1986.

JONQUERES Jean, *Droit d'auteur et droits voisins, la loi du 3 juillet 1985*, colloque de l'IRPI, Paris 21 et 22 novembre 1985, Librairies techniques, 1986.

KEREVER André, *Protection par le droit d'auteur ou protection "sui générés" ?* - In : Banques de données et droit d'auteur - colloque IRPI, Paris, 27 novembre 1986, Librairies Techniques.

LEMONNIER Béatrice, *L'originalité des logiciels, l'officialisation ou la nouveauté*, Mémoire de DESS, Université de Paris 2, 1988.

LAMBERT Bernard, *Nature juridique des programmes d'ordinateur*, thèse de doctorat en droit, Université de Paris 2, 1969.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

LAMBERTERIE (de) Isabelle (sous la direction de), *Le droit d'auteur aujourd'hui*, Editions du CNRS, Paris, 1991.

LATOUR Bruno, *La découverte est une invention*, in *Objectif Science*, WorldMédia/Libération n°3, Paris, décembre 1991.

LE STANC Christian, *La protection juridique des logiciels* - Cahiers du droit de l'entreprise, supplément n°24, 14/6/84.

LUCAS André, *La protection des créations industrielles abstraites*, Paris, Librairies Techniques, 1975.

LUCAS André, Jurisclasseur "*Propriété littéraire et artistique*", fascicule 303-1, n°36.

LUCAS André, *note sous Cass. civ. 1e*, 2 mai 1989, JCP 90, II, 21392.

LUCAS André, *Le droit de l'informatique*, Presses Universitaires de France, Collection Thémis, Paris, 1987.

MARIANI Joseph, *Les techniques d'entrée/sortie vocales*, Encyclopédie CIMAB, tome 5, Paris, 1988.

MARTIN Jean, *Quelques aspects des enjeux* - In : banques de données et droit d'auteur - Colloque IRPI, Paris, 27 novembre 1986, LITEC, 1987.

MAURY Jean-Pierre, *Comment la terre devint ronde*, Paris, Editions Gallimard, Collection La Découverte, 1989.

MEHL-MIGNOT Henri, *L'informatique au regard de la propriété intellectuelle : banques de données et droit d'auteur* - In : appropriation et circulation de l'information, Brises, CNRS, novembre 1982, n°12.

OCDE (ouv.col.), *Technologies de l'information et perspectives économiques*, OCDE, Collection ICCP n°12, Paris, France, 1987.

OCDE (ouv.col.), *Technologies de l'information et nouveaux domaines de croissance*, OCDE, Collection ICCP n°19, Paris, France, 1989.

PAUL Gérard, CATTENAT Annette, *Etude de quelques langages documentaires*, Université Paris VIII, 1977.

PIATTI Marie-Christine et GAUBIAC Yves, *La création artistique assistée par ordinateur, problèmes de droit d'auteur*, RIDA, Paris, 1983.

PIERREL Jean-Marie, *Dialogue oral homme-machine*, Hermes, Paris 1987.

PIOT Jean-Michel, *Les créations informatiques complexes* - Mémoire de D.E.A. , Montpellier, 1987.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

PLAISANT Robert, *La protection du logiciel par le droit d'auteur*, Gazette du Palais n° 83-1, 348.

PLAISANT Robert, *Jurisqueur propriété littéraire et artistique*, Fasc. 80, n°107.

POITEVIN Christine, *Un cadre juridique harmonisé pour les langues européennes*, In : Les industries de la langue, enjeux pour l'Europe - Actes du colloque de Tours, Revue Encrages, Paris, n° 16, 1986.

POUILLET Eugène, *Traité théorique et pratique de la propriété littéraire et artistique*, Editions Marchal et Billard, Paris, 1908.

REY Alain, *La terminologie, noms et notions*, Presses Universitaires de France, Collection Que sais-je ?, n°1780.

REY Alain, *Encyclopédies et dictionnaires*, Presses Universitaires de France, Collection Que sais-je ?, n°2000.

REY-DEBOVE Jacqueline, *La métalangue*, Editions Le Robert, Paris, 1978.

RISPAIL Jean-Luc, *Les surréalistes, une génération entre le rêve et l'action*, Gallimard, Paris, 1991.

SABAH Gérard, *L'intelligence artificielle et le langage*, Hermes, Paris 1988.

SALOMON Jean-Jacques, *Le gaulois, le cow-boy et le samouraï*, Etude CPE, ADITECH, Paris, 1985.

Silberztein, Max, *Dictionnaires électroniques et analyse lexicale automatique*, Thèse de doctorat, Université Paris 7-LADL, Paris, France, 1989.

SINCLAIR J.M., *Lexicographic evidence*, in Dictionaries Lexicography and Language Learning, R. Ilson (ed.), ELT Documents, Pergamon Press, 1985.

TRÖLLER Alois, *Urheberrecht und ontologie*, 50, UFITA, 1967.

ULMER Eugène, *La protection par le droit d'auteur des oeuvres scientifiques en général et des programmes d'ordinateur en particulier*, RIDA, octobre 1972, n°79.

VERNANT Jean-Pierre, *Les origines de la pensée grecque*, Paris, Editions Maspéro, 1990.

VIGNAL Jean-Christophe, *Systèmes experts et droit*, Cahiers Lamy, Paris, mai 1989 (C).

VIVANT Michel, LE STANC Christian, GUIBAL Michel et RAPP Lucien, *Lamy Droit de l'informatique*, Editions Lamy SA, Paris, 1992.

Contribution à une définition
du domaine des industries de la langue :
l'élucidation des statuts juridiques

VIVANT Michel, *Brèves réflexions sur le droit d'auteur suscitées par le problème de la protection des logiciels*, informatica e diritto, Florence, 1984.

VIVANT Michel, *Informatique et propriété intellectuelle*, JCP, 1984-1, 3169.

VIVANT Michel, *A propos des "biens informationnels"*, JCP, 1984-1, 3132.

VIVANT Michel, *Le logiciel au Pays des merveilles*, JCP, 1985-1, 3208.

VIVANT Michel et IRETIJ, *Éléments de réflexion pour une protection juridique des didacticiels*, IRETIJ, Université de Montpellier, 1985.

VIVANT Michel, *Les nouveaux biens*, Cahiers du droit de l'informatique, Editions Lamy, Paris, 1986 (H).

VIVANT Michel, *Intervention au colloque "Banques de données et droit d'auteur"*, Colloque IRPI, Paris 27 novembre 1986, Librairies Techniques, 1987.

VIVANT Michel, *Le "Produit informatique", discours sur un discours*, Paris, Dalloz, 1989, chr.p.140.

VIVANT Michel, *Le programme d'ordinateur au Pays des Muses*, JCP, Edition E, 1991-1, 94.

