

Premier forum

L'informatisation des langues

Président

Christian Fluhr
Professeur à l'INSTN-MIST

Rapporteur

Françoise Néel
LIMSI-CNRS

Intervenants

Roberto Cencioni
D.G. XIII-E4

Yves Normandin
*Centre de Recherche en informatique
de Montréal*

Stéphane Chaudiron
*Ministère français
de l'Enseignement supérieur et de la Recherche*

André Danzin
*Président du Forum international
des Sciences humaines*

Bernard Nomier
Société GSI Erli

Cheickh Tidiane Sow
Renault SA

Frédéric Doll
Machina Sapiens

Les programmes Ingénierie de la langue de la Commission de l'Union européenne

Je vais très rapidement préciser quelques aspects des activités communautaires liées au génie linguistique. J'appartiens moi-même à la Commission européenne, Direction générale 13, et depuis bientôt vingt ans, une unité s'occupe de questions liées à l'ingénierie du langage à Luxembourg et je suis responsable des activités qui s'inscrivent dans le cadre des programmes de recherche et de développement. Je ne veux pas m'attarder sur ces aspects des autoroutes de l'information. Il est évident que du fait du développement technologique multiple, nous allons très rapidement aller vers une société câblée sans doute et si tout se passe bien peut-être vers une société de l'information. Donc le risque est que dans cinq ou dix ans, nous nous trouvions dans un système qui ne reprendrait que de la ferraille et de la robinetterie, le défi est celui de tirer profit de cette évolution, pour rendre notre société et nos industries plus efficaces, plus démocratiques. J'essaie de montrer que, au fur et à mesure que notre société, nos industries deviennent tributaires de s connaissances, se pose, en Europe, mais partout dans le monde, un problème de contenu et d'accès à ces connaissances.

En Europe, il est important que le contenu véhiculé par les autoroutes de l'information soit à contenu européen et que nos sociétés aient un accès qui, en plus des attributs habituels, des priorités, interfaces et autres, va comporter un volet multilingue. Et nos efforts à la Commission sont de convaincre les autorités politiques que la maîtrise des langues naturelles est un élément important pour arriver à un résultat considérable, à la fois le contenu et l'accès à l'information.

Comme je le disais je m'occupe des aspects technologiques ou de recherches industrielles, et nous n'ignorons pas les aspects politiques et culturels, des programmes auxquels nous travaillons. Bien évidemment, il y a un problème de compétitivité industrielle, de croissance, de qualité de vie, de qualité des services publics rendus au public, mais il est important de souligner que des activités dans le génie linguistique, le

multilinguisme ont un impact direct sur l'intégration communautaire, pour la cohésion de nos sociétés. Elles doivent permettre entre autres d'assurer un développement harmonieux à différents groupes et nations, au delà des États, et il y a un problème culturel que d'autres intervenants ont sans doute souligné. Nous sommes conscients de cela au point où, en parallèle avec la réalisation du plan-cadre pour le génie linguistique, ma direction travaille en ce moment à un projet, une proposition de programme qui viserait à promouvoir, à stimuler une industrie multimédia liée au contenu européen de l'information, et doit remettre dans quelques semaines au Conseil, au Parlement une communication qui montre l'impact de la société de l'information sous l'angle linguistique et qui prévoit un nombre de mesures additionnelles d'accompagnement qui viendront s'ajouter aux activités actuelles de recherche et de développement.

Vous n'êtes pas sans savoir que le plan-cadre comprend une vingtaine de programmes spécifiques de recherche et de développement dont trois s'attachent aux questions de technologie de l'information des communications.

Le plan télématique au sein duquel s'inscrit le génie linguistique, se veut résolument tourné vers l'utilisateur du marché, et de ce fait, il se place en aval d'autres programmes, tels que ce qui était anciennement le Plan Esprit AS. Vous avez l'esprit informatique et technologies de l'information et réseaux et systèmes de télécommunication. Le plan télématique est un programme de validation, d'intégration qui vise des solutions dans un contexte environnement utilisateur.

Les trois programmes représentent globalement un investissement à l'échelle européenne de deux milliards d'écus, de l'ordre de 13 milliards de francs. La mission du plan télématique, est de contribuer à l'efficacité de la compétitivité des industriels, l'efficacité la qualité du travail également des services, par l'utilisation des moyens technologiques. Donc ce n'est pas un programme technologique ; il s'agit d'assurer que cela se traduise par des processus innovatifs pour les utilisateurs, par la prise en compte de ces technologies par les marchés, par les entreprises, par la société.

Donc nous demandons à nos projets de stimuler l'innovation, d'encourager le dialogue entre chercheurs, développeurs et utilisateurs. Nous insistons bien sur la viabilité, la pérenité des résultats ainsi obtenus sur le marché au sein de la société. Nous espérons que nos activités permettront ainsi de susciter, d'accélérer les investissements privés, nationaux publics-nationaux, qui vont tous aller dans la même direction c'est-à-dire la société de l'information.

Le plan télématique lui-même est un programme qui dispose d'un budget d'environ 850 millions d'écus, sur plus de quatre ans, qui comporte deux volets : des activités sectorielles liées à des domaines d'activités verticaux, des services publics ou d'intérêt public, tels que les administrations nationales ou territoriales, (enseignement, bibliothèques, etc) ; Et puis des secteurs transversaux qui intéressent l'ensemble de ces secteurs et puis d'autres

domaines d'activités professionnelles, économiques, industrielles, y compris, bien évidemment, le secteur privé, qui sont l'ingénierie linguistique de l'information télématique. Environ 80 millions d'écus ont été affectés aux recherches et développement en matière de génie linguistique pour la période en question. Cela représente une croissance assez importante par rapport au troisième plan-cadre, où nous ne disposerions que de 25 millions d'écus. C'est encore très largement insuffisant.

Alors parmi tous les thèmes possibles, imaginables nous en avons sélectionné trois pour les quatre ans à venir, qui sont :

- assurer la disponibilité de systèmes de gestion documentaires plus performants,
- assurer un accès et un accès sélectif plus efficace aux informations
- garantir une meilleure interaction interpersonnelle ou entre l'homme et la machine, par une communication plus naturelle, à caractère nouveau interpersonnel ou commercial ou professionnel.

Ce sont les trois axes. Ces faits sont bien connus. 40 % du temps consacré par les cols blancs au travail tourne autour de la gestion documentaire, 1 milliard de messages électroniques sont échangés chaque mois sur Internet, ce qui conduit à ce que les Américains appellent l'*information mass*, une masse énorme d'informations, dont il est pratiquement quasiment impossible de retirer quoi que ce soit, et un quart de la population américaine est incapable de se servir d'un clavier, alors que 40 % des PC vendus aux États Unis le sont à des fins domestiques, donc là il faut des paradigmes des moyens d'accès d'opération nouveaux.

La portée de nos activités : d'une manière très sommaire, elle est bien évidemment plus large que la traduction et s'attache à des thèmes tels que la rédaction de documents techniques, le classement et la recherche documentaire, et plus généralement la reconnaissance, la synthèse vocale, l'apprentissage de langues étrangères. Nous demandons à nos projets d'adopter une approche globale, c'est-à-dire poser les problèmes et proposer des solutions à ces problèmes. De ce fait, nos projets vont exploiter les technologies informatiques, de télécommunication, ils vont devoir consacrer une partie de leurs efforts aux études de marchés, à l'analyse et la modélisation des besoins des utilisateurs et processus de production, de fonctionnement de l'entreprise, l'ergonomie des systèmes, etc.

C'est donc une approche globale qui va au delà de ce qu'on comprend souvent, trop souvent, par le mot "génie linguistique", qui veut dire la traduction automatique.

Les objectifs fixés par le conseil sont d'améliorer la convivialité des systèmes télématiques, d'en assurer une acceptation plus large, plus rapide, et d'un déploiement de ce même système au delà des frontières nationales. À côté de cela nous avons une ligne d'action qui vise à développer, à promouvoir le développement et la distribution des ressources langagières ou linguistiques utilisables, à vocation très large, dictionnaire

électronique, corpus de texte, base de données, terminologie multilingues, etc, et par des efforts modestes mais importants de recherches, assurer une présence significative en Europe dans les systèmes à composante linguistique.

Le secteur cible, contrairement à d'autres domaines d'activités télématiques qui sont liés à un secteur spécifique des administrations ou du secteur des transports, nous sommes un programme horizontal, donc transversal, donc nous touchons un peu à tout, aussi bien le secteur public que différentes branches industrielles et commerciales, que les professions libérales indépendantes. Et nous avons déjà, depuis 91, beaucoup de projets dans chacun de ces secteurs.

Quels sont les intervenants ? Contrairement à ce que certains imaginent, je dirais dans les milieux politiques, le génie linguistique n'est plus simplement l'affaire de petites équipes de chercheurs ou de *gourous* informatiques, ça devient de plus en plus l'affaire d'utilisateurs et d'industriels. Nous sommes passés, depuis quatre ou cinq ans, d'une situation où il n'y avait pratiquement aucun apport d'utilisateurs ou d'industriels en zone européenne, 0 % tout simplement ! à une situation dans laquelle entre 40 et 60 % des ressources déployées dans l'espace communautaire viennent du secteur public ou bien d'organismes qui se placent en tant qu'utilisateurs du génie linguistique. Ce ne sont pas des gens qui veulent développer des nouveaux processus technologiques, ce sont des utilisateurs. Et si nous regardons ce qui se passe dans d'autres contextes européens, le programme *Esprit*, *Poste*, *Copernicus*, *Eurêka*, et bien d'autres, partout il y a des projets qui touchent directement ou indirectement aux questions de l'ingénierie linguistique. Dons, lorsque j'ai dit que mon équipe s'occupe en ce moment de quarante projets, et qu'avec d'autres collègues à Luxembourg, nous gérons soixante projets, cela ne représente que la moitié des activités communautaires en matière de génie linguistique.

Par rapport à nos activités passées, nous prenons une approche résolument globale et multidisciplinaire. Nous mettons l'accent de plus en plus sur les utilisateurs, sous les aspects liés à la validation, l'expérimentation sur le terrain en terme d'utilisateurs. Ce n'est plus une machine dans un coin avec un chercheur ou un ingénieur, il faut que ça se passe chez l'utilisateur. Il faut viser des solutions pratiques, concrètes. Il faut surtout s'attaquer aux dimensions multilingues et multimédias. Ça c'est donc le paradigme des banques communautaires. Donc il faut en arriver à maîtriser les techniques mais surtout s'assurer des quotas importants du marché multimédias qui est en train d'exploser très rapidement. 70 % des CD-ROM fabriqués de nos jours servent de supports pour des applications multimédias. Pour l'instant, en termes de chiffre d'affaires, ce n'est pas considérable, quoique ce soit tout à fait respectable, mais ça risque de croître très rapidement dans les années qui viennent.

Contrairement aux activités jusqu'à l'année dernière, du fait que nous disposons de moyens plus importants, nous pouvons aborder des applications pilotes dans le détail

nettement plus importantes, donc avec un impact beaucoup plus important. Des applications du traitement de la parole, applications vocales : jusque là nous nous sommes un peu cantonnés dans le traitement de l'écrit, et nous passons de la phase d'études, définition expérimentation, à la phase de construction de bases de données linguistiques, avec un effort qui ne peut être que complémentaire aux actions en cours au plan national.

Quelles sont donc les lignes d'action de ces programmes de génie linguistique ? Elles sont au nombre de trois. Applications pilotes, tout cela à l'air familier maintenant, création de gestion de documents, accès aux services d'informations et de communication, traduction assistée automatique et semi-automatique interactive, apprentissage de langues étrangères. Application pilote signifie réalisation avec et chez le client. Deuxième volet très important dans l'avenir, ressources utilisables, là se pose le problème de la création dans certaines langues, dans certains pays, très souvent, toujours le problème d'assurer un format et des normes des formes d'échanges, et malheureusement trop souvent, un problème de diffusion, de dissémination, de réutilisation même des ressources existantes. Ceci pose toute une série de questions d'ordre juridique auxquelles on réfléchit aussi bien à la Commission que dans les milieux français.

Troisième volet, peut-être modeste en termes budgétaires, mais très important, la recherche en génie linguistique ; N'oubliez surtout pas qu'il y a d'autres actions telles que le programme *Dictée* qui consacre une partie importante de leur budget (jusqu'à 5 millions d'écus), à la recherche de base. Le programme télématique n'est pas un programme voué à la recherche de base.

Nous avons lancé le premier appel à propositions le 15 décembre dernier, la date de clôture est le 15 mars, et il vise deux des trois lignes d'action, notamment les applications pilotes et la recherche. Un deuxième appel à propositions sortira au *Journal officiel de la Communauté* le 15 mars, il ne visera que les ressources et autres outils linguistiques réutilisables, à caractère général ; et enfin nous discutons mardi prochain à un troisième appel à propositions qui devrait être maintenant annoncé le 15 septembre prochain.

Un dernier mot pour ceux qui seraient intéressés par le volet recherche : dans le programme génie linguistique, nous avons englobé un très vaste champ de thèmes, donc la plupart des thèmes de recherche sont repris, mais nous insistons finalement sur trois dimensions : dimension industrielle, et je dirai potentiel d'exploitation des résultats de recherche, le fait que cette recherche puisse effectivement, concrètement contribuer à la société d'information ; Ce sont des thèmes dont l'essentiel est connu, information pour le citoyen, pour le client, le consommateur et l'entreprise, et qui visent des tâches, des environnements, des utilisateurs en contexte de travail, qui soient propres, qui apparaissent quelque part au sein des utilisateurs. Voilà pour le volet recherche. Pour le reste, au niveau des thèmes, il y a un éventail assez large, chacun trouvera un peu les thèmes qui lui sont propres. J'espère avoir donné une idée générale de nos activités.

Le programme américain ARPA-HLT : une expérience à élargir

Introduction

Il existe aux États-Unis, et ce depuis nombre d'années, un très important programme de recherche nommé HLT, pour « Human Language Technologies ». Le but de ce programme est de promouvoir la recherche et le développement dans le domaine du traitement automatique de la langue parlée et écrite. Le programme est financé à partir des fonds de l'agence américaine ARPA, c'est-à-dire « Advanced Research Project Agency ».

À une ou deux exceptions près – par exemple les « AT & T Bell Laboratories » – tous les laboratoires de recherche importants en traitement du langage humain doivent en grande partie leur existence aux fonds qu'ils ont reçus grâce à ce programme de recherche. On peut par exemple penser aux groupes à BBN, au Stanford Research Institute (SRI), au MIT, ou à l'université Carnegie Mellon (CMU), pour ne nommer que les plus connus.

Au-delà des fonds de recherche rendus pour les activités de recherche et développement en tant que telles, il existe une contribution tout aussi, sinon plus, importante de ce programme au développement technologique dans ce domaine. Cette contribution est la création d'une infrastructure de support au développement technologique, qui a jusqu'à présent été basée sur les principes suivants :

- Le ralliement des efforts autour des « tâches » communes visant à permettre le développement de certaines technologies qui, tout en étant aussi génériques que possibles, ont été identifiées comme stratégiques. En plus d'orienter les efforts dans certaines directions stratégiques, l'existence de tâches communes facilite grandement la comparaison entre différents systèmes, permettant ainsi d'en arriver à une meilleure appréciation des technologies sous-jacentes.
- La création de corpus de voix et de texte, dont l'objectif est d'abord et avant tout

d'offrir la matière première nécessaire aux développements technologiques accomplis dans le cadre des « tâches » identifiées. À cette fin, un nouvel organisme, le « Linguistic Data Consortium » a été créé en 1992 afin de coordonner les activités de production de corpus et d'en assurer la distribution.

- Le développement et le partage d'outils d'intérêt commun (outils de normalisation de transcriptions, d'apprentissage de modèles du langage, lexiques, etc.)
- L'évaluation formelle des technologies développées dans le cadre du programme, lors de « concours » au cours desquels les systèmes développés dans différents laboratoires sont testés sur des nouvelles données. Ces évaluations sont coordonnées par le National Institut of Standards and Technologies (NIST), qui est l'organisme du gouvernement américain en charge des standards technologiques.
- Le partage des résultats de recherche lors de colloques au cours desquels les résultats des dernières évaluations sont présentés et discutés.

Le programme HLT

Il est indéniable que le programme HLT a eu un impact considérable sur le développement des technologies de traitement du langage parlé et écrit. En fait, si on remonte beaucoup plus loin dans le passé, on se rend compte que la recherche dans le domaine de la reconnaissance de la parole et à toutes fins pratiques été lancée à la fin des années 60 par l'ancêtre du programme actuel. L'organisme subventionnaire qui à l'époque s'appelait ARPA, a par la suite changé son nom pour DARPA (le « D » voulant dire « défense ») pour finalement revenir au nom ARPA il y a deux ans.

Les résultats de ce premier effort de recherche important, d'une durée de cinq ans, ont constitué à l'époque une profonde déception pour plusieurs. Comme dans bien des branches de l'intelligence artificielle, on avait grossièrement sous-estimé la difficulté du problème. Par contre, si on y regarde de plus près, ce premier programme de recherche a eu des impacts positifs importants. Par exemple :

- Ce programme nous a permis de réaliser que les problèmes reliés au traitement automatique du langage humain étaient beaucoup plus complexes et difficiles que l'on se l'imaginait.
- Plusieurs des techniques que l'on utilise actuellement en reconnaissance de la parole (dont les modèles de Markov et le « Viterbi beam search ») ont été initialement développées dans le cadre de ce programme.
- Plusieurs des meilleurs laboratoires de recherche actuels aux États-Unis (dont IBM, BBN, ou l'Université Carnegie Mellon) ont vu le jour grâce à ce programme.

Dans sa forme actuelle, le programme ARPA (ou plutôt DARPA) semble avoir vu le jour au milieu des années 80. Je dis « semble » parce que, comme il ne semble pas exister

d'historique écrit de ce programme, je dois me baser sur la mémoire collective des gens impliqués. Le modèle utilisé à l'intérieur du programme est le suivant :

1. Définir une « tâche » qui, tout en étant réaliste, est suffisamment difficile pour permettre de travailler sur des problèmes de recherche fondamentaux en traitement automatique du langage parlé et écrit.

2. Produire le matériel de base nécessaire aux travaux de recherche, tels les corpora de voix et de texte, de même que les outils nécessaires à leur manipulation.

3. Faire des « tests », à intervalles réguliers, des différents systèmes produits par les laboratoires sous contrat de ARPA dans le but de comparer la performance avec différents types d'approches. Ce n'est que récemment que des laboratoires « volontaires » (par exemple AT & T, le LIMSI, Cambridge University et le CRIM) ont été invités à participer à ces tests.

Les tâches du programme

La première tâche de la série s'appelait « Resource Management ». Ses caractéristiques étaient :

- reconnaissance d'un vocabulaire de 997 mots en parole continue (phrases lues), dépendant et indépendant du locuteur
- modèle du langage de type « word-pair » avec une perplexité de 60
- microphone près de la bouche et de bonne qualité, faible bruit ambiant.

Au départ les premiers résultats obtenus avec cette tâche ont été assez médiocres. Cependant, en quelques années, l'amélioration de la performance des systèmes évalués a été telle qu'il était temps de passer à des tâches plus difficiles.

Il est important de remarquer que, aussi simple soit-elle cette tâche a eu un impact considérable sur le domaine de la reconnaissance de la parole. D'abord parce qu'elle a démontré la faisibilité de reconnaître de tels vocabulaires en parole continue (ceux qui sont du domaine se rappellent probablement que le système SPHINX, développé dans le cadre du programme DARPA, avait été annoncé à l'époque comme une percée majeure). Ensuite parce qu'elle a été à l'origine du développement de modèles de « phones en contexte » maintenant presque universellement utilisés pour la modélisation de grands vocabulaires ou de vocabulaires flexibles.

La tâche Ressource Management a été suivie par deux nouvelles tâches assez différentes. D'abord, la tâche ATIS (pour Air Travel Information System) a été créée dans le but de développer des techniques permettant l'interprétation de la langue naturelle. En effet, un système ATIS vise à permettre à un utilisateur d'obtenir, avec des requêtes verbales totalement spontanées, les informations nécessaires à la planification d'un voyage aérien. Par « requêtes spontanées », on veut dire que l'utilisateur est libre de s'exprimer comme il

le désire, et en particulier d'utiliser les mots de son choix. Dans le cas de la tâche ATIS, l'accent est mis non pas sur le pourcentage de reconnaissance en tant que tel, mais bien sur la capacité du système d'interpréter correctement la signification des requêtes du locuteur. Pour fins d'évaluation des systèmes, la façon dont cette capacité est mesurée est de comparer la réponse produite par le système à celle qui, après définition de la tâche, aurait dû être produite. L'approche utilisée par le système pour produire sa réponse n'a, dans cette évaluation, aucune importance. En particulier, il importe peu si le système tente de reconnaître tous les mots de la requête ou au contraire s'il se contente de seulement reconnaître les mots sémantiquement importants.

La seconde tâche, qui s'appelle simplement « Continuous Speech Recognition » (CSR) vise spécifiquement des applications de dictée grand vocabulaire en parole continue. Le corpus initial créé pour cette tâche était exclusivement constitué d'extraits du « Wall Street Journal ». Plus récemment, le corpus s'est enrichi de nouveaux journaux (dont le San Jose Mercury News) d'où son nom : NAB, pour « North American Business News ». La tâche CSR existe maintenant depuis quatre ans et, plutôt que de garder la tâche fixe comme c'était le cas auparavant, la définition des évaluations CSR a changé à chaque année, devenant chaque fois plus difficile et plus complexe.

Par exemple, la première année, le vocabulaire avait été fixé à 5000 mots « fermés » (c'est-à-dire que les phrases test étaient garanties de ne pas contenir les mots « hors vocabulaire »). Par la suite, le vocabulaire a été augmenté à 20000 mots « ouverts », pour finalement devenir l'an passé sans limites de vocabulaire. Pendant ce temps, on a greffé aux tests standards, jugés obligatoires, des tests satellites optionnels visant à attaquer un aspect spécifique du problème de reconnaissance grand vocabulaire tel que la reconnaissance en présence de bruit ambiant, sur le réseau téléphonique l'adaptation au locuteur, etc. Ce nouveau modèle d'évaluation a été surnommé le « hub and spokes paradigm », par analogies avec les termes correspondants utilisés dans le domaine du transport aérien pour représenter les aéroports pivots (« hubs ») et les autres (« spokes »).

Un bilan du programme

Afin de pouvoir tirer certaines leçons dont on pourrait tirer partie, je vais tenter de faire un bilan du programme HLT dans sa forme actuelle. C'est clair que la qualité du CRIM de participant volontaire et, de surcroît, de nouveau venu dans ce programme ne me permet pas d'en dégager une vision globale. Cependant, plusieurs conclusions peuvent tout de même être tirées sans trop de risques.

Au niveau des avantages du programme, ils sont indéniables. Il ne fait aucun doute que le programme HLT a énormément contribué à l'avancement des connaissances sur le traitement automatique du langage parlé et écrit. De plus, les corpus de voix et de texte

produits au fil des ans constituent une richesse – la matière première même dans la recherche dans le domaine – que la communauté de recherche pourra encore utiliser de nombreuses années. Malheureusement, bien que le LDC commence depuis peu à s'intéresser sérieusement aux aspects multilingues, l'essentiel de cette richesse est encore en anglais américain. C'est d'ailleurs ce qui a poussé plusieurs laboratoires non américains comme le LIMSI, à l'université de Cambridge, Phillips et le CRIM, à joindre en quelque sorte le programme et à travailler sur une langue qui n'est pas la leur.

Le programme a aussi eu l'effet de contribuer au démarrage de plusieurs laboratoires de recherche dans le domaine aux États-Unis. On assiste d'ailleurs actuellement à une explosion du nombre d'applications commerciales, pour lesquelles une grande partie du crédit doit aller directement ou indirectement à ARPA. Deux des laboratoires principaux supportés par ARPA, BBN et SRI, ont d'ailleurs récemment chacun créé une nouvelle division spécialement dans le but de commercialiser des technologies.

Du point de vue de la recherche, plusieurs développements importants ont été directement issus du programme HLT. On peut par exemple penser à la modélisation de « phones en contexte » ou encore aux récentes stratégies d'accès lexical développées dans le cadre des travaux sur la dictée grand vocabulaire.

D'un autre côté, le programme HLT a aussi ses problèmes. D'une part les administrateurs des fonds semblent reprocher le peu d'interaction entre cette communauté de chercheur et la communauté de recherche en interfaces personnes – systèmes. On semble aussi trouver que la technologie ne constitue qu'un des éléments importants d'une application à succès et qu'on n'a peut-être pas suffisamment consacré d'efforts aux autres éléments.

Il y a aussi eu des problèmes concernant la propriété intellectuelle. Par exemple, lors du colloque ARPA de 1993, BBN a démontré son système de dictée temps – réel avec un vocabulaire de 20 000 mots. Ils ont cependant refusé d'en expliquer le fonctionnement sous prétexte que ce travail avait non pas été payé avec l'argent de ARPA, mais plutôt avec celui de BBN, ce qui avait alors provoqué des débats assez animés.

Aussi, l'évaluation, cet outil si important lorsqu'on veut mesurer les avantages respectifs de différentes approches, méthodes, ou technologies, peut aussi devenir un frein à l'innovation. En effet, si on met un accent trop prononcé sur les résultats obtenus, on encourage tous les laboratoires à adopter immédiatement les approches ayant le mieux fonctionné lors de la dernière évaluation et, similairement, on décourage l'emprunt de pistes apparaissant trop risquées. Le problème d'évaluation peut aussi devenir un véritable casse-tête aussitôt que l'on aborde les aspects d'interaction entre le système et l'utilisateur.

Mais, de façon globale, le programme a eu tellement de succès qu'on se demande maintenant s'il a toujours sa raison d'être (à tout le moins dans sa forme actuelle). En effet, avec l'explosion du nombre de débouchés commerciaux possibles, il y a maintenant

beaucoup d'argent privé investi dans le domaine aux États-Unis, ce qui remet en question le besoin pour ARPA de continuer à supporter financièrement des laboratoires. Ceci est d'autant plus vrai que, au cours des trois dernières évaluations, les meilleurs laboratoires n'étaient pas toujours ceux supportés par ARPA. Il n'est donc pas impossible que, maintenant que plusieurs laboratoires peuvent voler de leurs propres ailes, ARPA limite son support aux aspects mobilisateurs du programme, c'est-à-dire la création de corpus, de même que l'organisation d'évaluations et de colloques.

Conclusion

Il est ironique de voir que, au moment où la Francophonie parle de mettre en place une infrastructure nécessaire à la recherche et au développement en technologies de traitement du langage humain, certaines rumeurs circulent comme quoi le programme américain correspondant pourrait subir des coupures importantes.

La situation est cependant bien différente. Alors qu'il existe pour l'anglais une richesse inestimable en termes de ressources linguistiques, il n'existe tout simplement rien d'équivalent pour le français. Les seuls corpora importants dont on connaît l'existence sont en général impossible à obtenir, généralement pour des raisons de propriété intellectuelle.

Il est vrai que plusieurs des progrès technologiques accomplis sur l'anglais vont aussi s'appliquer à d'autres langues, incluant le français. Cependant, chaque langue a aussi un très grand nombre de particularités, que ce soit sur le plan phonétique, prosodique ou grammatical, que l'on doit bien comprendre avant de pouvoir développer des systèmes compétents pour cette langue. Pour en arriver à cette compréhension, le matériel de base sont les corpora parlés et écrits, sur lesquels les travaux de recherche se feront. Encore faut-il qu'ils soient disponibles, cependant.

L'expérience du programme SLT de ARPA montre à quel point les progrès peuvent être rapides lorsque l'on réussit à coordonner les meilleurs laboratoires de recherche en les faisant travailler sur des problèmes difficiles mais communs à tous et en leur fournissant le matériel de base. Grâce à ce programme, l'anglais a, encore une fois, pris une longueur d'avance sur les autres langues. Les enjeux économiques et culturels sont trop importants pour que la Francophonie ne fasse pas tout en son pouvoir pour rattraper le temps perdu.

À cet égard, l'expérience des américains avec le programme HLT est très précieuse. Il ne faut pas nécessairement chercher à imiter ce programme en tous points, mais il est essentiel de chercher à s'en inspirer.

Les projets innovants en ingénierie linguistique

1. L'informatisation des langues : un enjeu stratégique

Les enjeux de l'informatisation des langues sont étroitement liés au formidable essor des technologies de l'information et de la communication dont le phénomène « autoroutes de l'information » est le dernier exemple. La mondialisation des échanges et la dématérialisation de l'information due à la révolution numérique risquent donc de modifier irrémédiablement le statut des différentes langues nationales dans le monde et soulignent fortement les enjeux culturels, industriels et économiques liés à la mise en place d'une véritable politique en faveur d'une ingénierie linguistique francophone dans le cadre du plurilinguisme.

Avant de présenter les actions menées par le ministère chargé de la recherche dans le domaine de l'ingénierie linguistique, je rappellerai brièvement l'importance de prendre en compte les dimensions économique et culturelle de l'informatisation des langues.

Un enjeu important

L'importance économique du domaine des industries de la langue s'apprécie à trois niveaux : en premier lieu, il s'agit d'un secteur qui correspond à un marché en tant que tel. Une étude réalisée par l'OFIL en 1994 estimait la taille du marché des produits et services à 600 MF en France et 1,5 MdsF en Europe. Cette étude, ainsi que plusieurs autres (Ovum, Frost et Sullivan et Bossard notamment), souligne par ailleurs le taux de croissance exceptionnel des différents segments du marché dont on peut établir la typologie suivante en terme de systèmes opérationnels :

- les systèmes d'indexation et d'interrogation de bases de données en langage naturel,
- les interfaces vocales et multimodales,
- les systèmes d'aide à la rédaction (vérification, correction et contrôle),

- les systèmes d'analyse de contenu, de filtrage et de routage de messages,
- les systèmes d'aide à la traduction et de traduction automatique,
- systèmes de génération de textes (écrits et oraux),
- les systèmes de dictée vocale (les dictaphones).

Il s'agit en deuxième lieu d'un marché dérivé, c'est-à-dire que l'utilisation de technologies linguistiques dans les systèmes ou produits d'information, qu'il s'agisse d'un système complet de GED ou d'un simple traitement de textes, peut améliorer les performances globales du système d'information. Dans le chiffre d'affaires global généré par la vente de ces systèmes, il est donc nécessaire d'identifier la part qui relève de la présence des modèles linguistiques afin de cerner plus précisément le marché des industries de la langue. De plus, l'intégration de modules linguistiques constitue de plus en plus un argument de vente décisif auprès des utilisateurs potentiels et, en ce sens, il conviendrait de conduire une étude très précise auprès des utilisateurs afin de mieux cerner les critères de choix pris en compte lors de l'achat d'un système.

Enfin, l'introduction des logiciels linguistiques dans les systèmes de traitement de l'information permet d'envisager des gains de productivité importants pour les entreprises. Dès à présent, dans les domaines de l'édition et de la gestion de documents et de la traduction technique, l'amélioration quantitative et qualitative des tâches est réelle. Dans les domaines de la messagerie électronique, de la gestion de la documentation ou des télécommunications en temps réel, des gains considérables de productivité sont également à attendre dans les prochaines années.

Un impact culturel immédiat

Compte tenu du développement extrêmement rapide des technologies de l'information et de la communication dans toutes les sphères de la vie professionnelle et privée, la disponibilité des outils de traitement de l'écrit et de la parole est absolument déterminante pour l'avenir des langues à moyen terme. D'ici la fin du siècle, les langues qui ne seront pas informatisées c'est-à-dire pour lesquelles on ne disposera pas de logiciels de traitement automatique du langage naturel, ne pourront plus prétendre à être des langues véhiculaires dans les domaines des sciences, des techniques et des affaires. Cette dimension culturelle des industries de la langue nécessite en particulier de veiller à la disponibilité des ressources linguistiques de base (terminologies, dictionnaires électroniques, grammaires) permettant la réalisation d'applications industrielles.

2. Les actions des pouvoirs publics

Les actions menées par les pouvoirs publics correspondent donc à cette volonté d'intégrer ces deux dimensions apparemment contradictoires. Elles s'articulent autour de

quatre axes d'intervention pour lesquels nous ne mentionnons que quelques projets significatifs à titre d'exemple (ne sont indiqués que les noms des chefs de projet) :

Le soutien au développement de ressources linguistiques

Objectif : rendre disponible (production et diffusion) les ressources linguistiques génériques afin de baisser les coûts de réalisation des logiciels de traitement de la langue sur le plan industriel et favoriser la recherche académique dans le domaine.

- GENELEX [Sema Group] : ce projet (1990-1995) vise à assurer la réutilisabilité des dictionnaires utilisés dans les applications d'ingénierie linguistique en définissant un format générique pour la description des ressources lexicales en français, italien, espagnol et portugais.

[Projet Eurêka]

- GRAAL [GSI Eri] : le but de ce projet (1992-1995) est d'offrir un ensemble d'outils en langage naturel (une boîte à outils) permettant le développement d'applications à moindre coût dans des domaines tels que l'interrogation de bases de données, l'indexation automatique ou la traduction assistée par ordinateur. Les langues concernées sont l'anglais, le français, l'espagnol, l'italien et le portugais.

[Projet Eurêka]

- Terminologies spécialisées : 12 projets retenus
[Appel à propositions *Terminologie 1991*-MESR-DISTB & DGLF]

Le soutien au développement d'applications pilotes

Objectif : démontrer la faisabilité des approches novatrices, assurer le transfert technologique des résultats de la recherche vers l'industrie et créer un contexte technologique favorable à une diversification de l'offre.

- Ante-Serveur intelligent [Sté Triel] : Réalisation d'une maquette d'un ante-serveur documentaire intelligent pour le grand public doté de fonctions assurant une aide à la formulation de la demande de l'utilisateur (choix des bases pertinentes, reformulation de la requête).

[Appel à propositions *Interfaces intelligentes* – Min. de la Recherche et Min. de l'Industrie – 1990]

- MAXIM'S [Sté SDC Informatique] : Constitution d'une méthode d'élaboration d'interface de base de données permettant un dialogue coopératif entre l'utilisateur et la base, et réalisation d'un prototype.

[Appel à propositions *Interfaces intelligentes* – Min. de la Recherche et Min. de l'Industrie – 1990]

- RIVAGE [Un. de Nancy-CRIN] : Réalisation d'une maquette d'interface interactive pour la recherche de données multimédia basé sur un processus d'apprentissage.

[Appel à propositions *Interfaces intelligentes* – Min. de la Recherche et Min. de l'Industrie – 1990]

- SERAPHIN [Sté Ediat et Un. Paris IV] : Réalisation d'une maquette de système de résumé automatique utilisant un système expert de repérage des phrases importantes basé sur la technique de l'exploration contextuelle.

[Appel à propositions *Ingénierie linguistique 1993* – Min. de la Recherche]

- SYRTE [Sté STEP Informatique] : Réalisation d'un prototype de système automatique de résumés de textes juridiques.

[Appel à propositions *Ingénierie linguistique 1993* – Min. de la Recherche]

- LOGOGRAPHIE [Sté Résoudre] : Réalisation d'une maquette permettant la génération automatique de résumés pour des applications en veille technologique. Intégration du module dans le logiciel RL-DOC.

[Appel à propositions *Ingénierie linguistique 1993* – Min. de la Recherche]

- ANTARES [Sté Bertin & Cie et Sté Madicia] : Spécification d'une plate-forme logicielle adaptée à la veille technologique basée sur des techniques de génération automatique de résumé de textes techniques.

[Appel à propositions *Ingénierie linguistique 1993* – Min. de la Recherche]

- EMIR [Sté T-GID] : Extension du projet Esprit Emir à la langue russe. L'objectif est de réaliser un système d'indexation automatique en texte libre et d'interrogation multilingue des bases de données textuelles.

[Min. de la recherche – 1994]

Le soutien à l'offre de produits et services

Objectif : faciliter l'intégration des technologies linguistiques dans les systèmes et produits d'information

- ILLICO [Sté PrologIA et CNRS] : Réalisation d'un environnement logiciel permettant de développer des systèmes pour analyser et synthétiser des phrases. Parmi les principaux domaines d'application, on peut mentionner les interfaces en langage naturel, les systèmes d'aide à la communication pour les personnes handicapées, les systèmes d'apprentissage de langues.

[Appel à propositions *Interfaces intelligentes* – Min. de la Recherche et Min. de l'Industrie – 1990]

- PRS [Sté Questel et Sté Kaos] : Conception et réalisation d'un module de correction orthographique en vue d'améliorer l'ergonomie des interfaces utilisateurs pour l'interrogation de banques de données sur les entreprises.
- EUROLANG [Sté Site] : le but de ce projet (1991-1994) est de développer un poste de travail du traducteur comprenant un système de traduction assistée par ordinateur fonctionnant pour l'anglais, l'allemand, le français, l'italien et l'espagnol.

[Projet Eurêka]

- CAROLUS [Sté Cora] : ce projet (1993-1996) vise à développer un système de Gestion Électronique de Documents (saisie, correction, indexation, classement et traduction des documents) pour le français et l'espagnol.

[Projet Eurêka]

- Fourniture de brevets [INPI & Sté Questel] : Réalisation d'un service vocal permettant la commande du texte intégral des brevets.

[Appel à propositions *Audiotext 1991* – Min. de la Recherche]

- AFP Sciences [AFP & Sté Cap Gemini Innovation] : Réalisation d'un service de consultation et de commande des dépêches scientifiques de l'AFP en texte intégral.

[Appel à propositions *Audiotext 1991* – Min. de la recherche]

- FLAUBERT [Sté Cora] : Réalisation d'un système général de génération de textes qui sera adapté au domaine de l'expérimentation chimique.

[Appel à propositions *Ingénierie linguistique 1993* – Min. de la Recherche]

Les actions d'accompagnement

Objectif : créer un contexte socio-technique favorable et rapprocher l'offre de la demande

Les actions d'information et de sensibilisation,

- colloques,
- préparation des actions de la DG XIII/E,
- réalisation d'études de veille technologique, de répertoires, catalogues...
- soutien à la revue *La Tribune des industries de la langue*, éditée par l'OFIL.

Les travaux de normalisation et de standardisation,

- groupe de travail sur la normalisation en terminologie [Un. de Rennes & AFNOR],
- groupe de travail sur la normalisation des technologies de l'information dans leurs aspects linguistiques.

L'évaluation des technologies et des produits.

En conclusion, je rappellerai que la liste des projets ci-dessus ne correspond pas à l'ensemble des projets qui ont été soutenus par les pouvoirs publics ces dernières années mais sert à illustrer les axes de notre programme. Rappelons également qu'il convient de se garder de tout triomphalisme et que, s'il est certain que des efforts importants ont été consentis pour favoriser l'informatisation de la langue française, la tâche qui reste à accomplir est encore immense avant d'aboutir à des systèmes industriels permettant un véritable traitement en profondeur du langage humain. Néanmoins, l'offre actuelle en

matière de produits et services d'ingénierie linguistique a atteint un niveau d'industrialisation (environnement de programmation, ressources génériques, modules réutilisables, début de standardisation...) qui permet une réelle intégration de ces technologies dans les systèmes d'information et les branches d'activité professionnelle les plus divers.

Le français sous le choc des inforoutes

1. La mode est aux autoroutes de la communication, encore appelées « inforoutes ». Les modes s'imposent comme des nécessités. On ne leur résiste pas davantage en techniques qu'en vêtements féminins ou en politique. L'affaire est née aux États-Unis où un groupe d'industriels en a vendu le projet au Vice-Président Al Gore dont le grand-père avait défini les objectifs Keynésiens du New Deal auprès du Président F. D. Roosevelt.

L'initiative américaine ne manque pas de caractéristiques colbertistes dont Al Gore est coutumier. En 1989, il avait publié « *Earth in the balance* » ; cet ouvrage écologique est un véritable hymne au dirigisme. Les intentions en matière d'infrastructure de l'information sont claires. Par des mesures gouvernementales d'anticipation, il s'agit d'introduire, en avance sur le marché, un stimulateur du développement des industries des technologies de l'information et des industries de la connaissance et des programmes.

Par cette voie, l'Amérique du Nord confortera sa supériorité dans les spécialités qui arbitreront, au début du prochain siècle, la puissance économique et la division internationale du travail. Le moment est particulièrement opportun, à la veille de l'ouverture des marchés mondiaux des télécommunications. Le marché européen, notamment, jusqu'ici protégé mais aussi parcellisé par les monopoles nationaux, va offrir, par la déréglementation, un terrain plein de promesses aux ambitions des multinationales américaines admirablement préparées à sa conquête.

Permettez-moi, au passage, de sourire des libéraux purs et durs, assez naïfs pour croire que la mondialisation du commerce s'accompagne partout d'un effacement du rôle de l'État.

2. La présentation théâtrale des inforoutes n'était pas, à vrai dire, nécessaire pour que le phénomène connaisse un essor considérable ; il est inscrit dans la force des choses, dans les conséquences sur la société des avancées technologiques.

Le succès d'Internet dont la vitesse de croissance au cours des deux dernières années a été prodigieuse, démontre que nous sommes en présence des conditions d'une sorte d'explosion critique grâce aux progrès accomplis sur la numérisation de l'information, la soif d'échanges de connaissances n'est plus limitée par les moyens de communication et de mémoire. Cette soif se satisfait et s'entretient comme un besoin inextinguible.

En fait, le monde est déjà doté, par les satellites et les câbles sous-marins, par les réseaux à large bande et par le téléphone d'autoroutes, de routes et de chemins vicinaux.

La circulation s'y accroît sans cesse. Le mouvement est en lui-même, moteur de nouvelles extensions. Indépendamment de toute intervention gouvernementale, nous sommes déjà dans la partie exponentielle de la logistique de croissance. Tous les décideurs économiques et politiques auraient du en être conscients depuis longtemps. Le coup de tonnerre américain a l'avantage de réveiller les esprits européens quelque peu endormis au moment même où l'exécution du programme AL GORE connaît ses premières difficultés administratives et financières. Rien n'est donc perdu pour notre continent.

3. Nous pouvons gagner si nous ne commettons pas des fautes de jugement. Il faut admettre – ce qui ne semble pas aller de soi – que le problème des infrastructures physiques de communication et de traitement est second par rapport au problème des informations proprement dites et des programmes. Le contenu prime sur le contenant. Des inforoutes en Europe pour véhiculer quoi ? et dans quelle langue ? Là sont les questions.

Bien que je ressente quelques craintes dominées par les ingénieurs, je commence à être rassuré sur la réceptivité du slogan « le contenu avant le contenant ». Le livre blanc de la Commission Européenne y insiste. Le message a été approuvé par le Conseil des Chefs d'État. Des programmes expérimentaux vont être mis en place, notamment par le Gouvernement de notre pays et par de puissants acteurs industriels. Ces signes sont encourageants. J'ai toutefois de sérieuses inquiétudes sur la manière dont les pays membres de l'Union Européenne, y compris les plus grands, vont pouvoir vaincre l'obstacle des effets d'échelle. L'identité européenne est, en effet, caractérisée par la variété. Cette diversité est une richesse qui doit être préservée mais elle s'oppose à l'anéantissement massif des frais de premier établissement des produits sur un large marché intérieur. Or dans l'économie des activités immatérielles où nous entraînent les inforoutes, les coûts sont presque totalement exposés dans la phase qui précède l'introduction sur le marché. Hypertextes, multimédias, hypermédias, programmes éducationnels, culturels, télévisuels sont affaires de conception et de réalisation de détails pour l'achèvement du prototype après quoi les coûts de reproduction ou de mises à disposition sont presque négligeables. Or quelle est, en Europe, la source de dispersion, de parcellisation fondamentale ? A l'évidence, c'est la diversité linguistique et la diversité culturelle qui accompagne l'usage de nos langues maternelles.

4. La problématique linguistique ressurgit ainsi comme un élément central de notre analyse. A première vue, la partie est désespérée. Dans les produits des industries de l'immatériel, les langues fractionnent le marché de l'Union européenne plus efficacement que toutes les dispositions du protectionnisme tarifaire, paratarifaire ou réglementaire que nous avons pu connaître pour le contrôle des biens et des services aux frontières. Pour les nouveaux médias et pour les programmes, le Marché Unique serait une fiction : le phénomène linguistique fait obstacle aux franchissements des seuils de compétitivité.

Cette conclusion pessimiste est probablement fausse. Car cette boucle réactive perverse en provenance de la langue peut se retourner en une contre – boucle vertueuse sur la langue. Je veux dire que la technique informatique qui semble nous incliner vers l'usage unique de l'anglo américain, nous propose aussi de merveilleux outils pour automatiser nos langues maternelles au point que leur diversité n'aurait plus d'effets négatifs.

Nous assistons en effet, à la naissance et au développement, encore hésitant, il est vrai, des industries de la langue capables de nous fournir des outils prodigieusement efficaces pour traiter nos langues.

Chacun d'entre nous, dans cette assemblée, en connaît les promesses et les insuffisances. Est – il nécessaire de les évoquer ? Synthèse et reconnaissance de la parole, identification de l'écriture manuscrite, machines à lire et machines à dicter, perfectionnement du traitement de texte, correcteurs d'erreurs sémantiques et syntaxiques, analyseurs de style, aides à la rédaction de textes contrôlés, aide à la traduction, systèmes d'éditions automatisés, aides à l'élaboration des hypertextes et des hypermédias, archivage et recherche de commentaires automatiques.

Je m'arrête là et vous renvoie à l'excellent « guide des produits et des services d'ingénierie linguistique » que vient de publier l'OFIL à l'initiative de l'AUELF – UREF.

Tous les experts sont d'accord. Sous l'impact de nouvelles technologies du génie linguistique, cette mutation du statut de la communication, est plus profonde, plus dense, plus riche de conséquences économiques et sociales que l'apparition de l'imprimerie. Elle nous fait passer de la civilisation de l'écrit à la civilisation de l'écran. Au lieu de s'étendre sur plusieurs siècles, le phénomène s'accomplit en quelques décennies.

5. En présence de cette métamorphose, il faut prendre conscience de la dimension de la Francophonie dans le monde. En projection vers l'année 2020, le français ne sera la langue maternelle que de 1 % des humains. On doit en outre avoir l'ambition que notre langue continue d'être le premier outil de la communication d'environ 3 % du reste de la population mondiale pour sa marche vers le développement.

Cette place modeste ne nous condamne pas à l'exclure. Si le français pouvait se situer à la pointe des perfectionnements fondamentaux apportés à l'usage des langues de

l'informatisation, l'obstacle des dimensions cesserait de jouer un rôle majeur tant les opérations de traitement de conversions, d'élaboration des gisements de connaissances et d'accès aux sources deviendraient peu onéreux.

6. A partir du moment où l'on sait que l'un des facteurs déterminants de la permanence du français comme langue de communication internationale est d'ordre technique, il devient facile de concevoir une stratégie offensive. La France est particulièrement bien placée pour concevoir cette stratégie sur le modèle de ses réussites antérieures, sur le modèle du nucléaire, de l'aérospatial, du minitel, du TGV ou des logiciels informatiques où elle a su exceller. Il suffit par un organe spécialisé, de mettre en œuvre la synergie de huit facteurs de succès aujourd'hui, sans liaison efficaces entre eux : la recherche fondamentale et appliquée, la constitution des ressources langagières, le développement industriel et commercial, la veille stratégique, la formation des utilisateurs et des spécialistes, l'évolution du droit sur la propriété intellectuelle appliquée aux traitements des langues, les alliances internationales et la protection du génie de la langue. Les conditions de succès de cette synergie sont définies dans le rapport intitulé « le français soumis au choc des technologies de l'information ; proposition pour une pacifique offensive » que j'ai eu l'honneur de présenter le 7 février dernier au Conseil Supérieur de la Langue Française. J'ai tout lieu de croire qu'une décision sera prise en faveur de la création d'un point focal actif pour le rassemblement de cette politique.

Du bon usage du dictionnaire électronique : des applications pratiques

Je représente ici une entreprise privée qui vit presque depuis vingt ans en faisant des applications de traitement automatique du langage naturel. Nous sommes une entreprise qui aujourd'hui fait une soixantaine de personnes et qui regroupe des informaticiens et des linguistes pour faire des projets industriels pour un certain nombre de clients et des projets de recherche et développement en ingénierie linguistique.

Avant de parler d'ingénierie linguistique proprement dite, et de types de projet que nous faisons, je voudrais replacer ça dans un cadre un petit peu plus général en posant une question pour savoir s'il y a une différence entre donnée et information.

Des données, il y en a un peu partout, il y en a dans des bases de données, il y en a dans des documents divers et variés, il y en a des quantités dans les entreprises. La question est : est – ce – que, si on a beaucoup de données, on a pour autant beaucoup d'informations ? Et en fait, le fait d'avoir beaucoup de données n'est pas pour autant une clé pour une bonne information ; une information c'est disposer d'une bonne donnée, sous la bonne forme, au bon moment, au moment où l'on en a besoin et pour la bonne personne ; sinon on a des gisements de données qui ne sont pas pour autant intéressants.

C'est une banalité maintenant que de dire que l'on a trop de données. Les bases de données deviennent gigantesques, les réseaux permettent de se connecter à des réservoirs énormes à l'intérieur ou à l'extérieur de l'entreprise et, en fait, on est un peu dans la situation de quelqu'un dans un désert qui est la proie d'un mirage, qui voit de l'eau partout mais qui n'a rien à boire. Et là c'est un peu la même chose, on voit des données partout, mais c'est de plus en plus difficile d'avoir de l'information parce que c'est de plus en plus difficile de trouver la bonne information pour la bonne personne, au bon moment, dans cette masse gigantesque de données. À tel point que je me demande si on ne peut pas se poser la question de savoir à propos des autoroutes de l'information dont on parle tant, si ce ne sont pas en fait, des autoroutes de données qu'on veut construire. Je

suis parfaitement d'accord avec ce que vient de dire Monsieur Danzin : « *Ce qui est important ce n'est pas le tuyau mais le contenu* » mais moi je pense que ce qu'il y a d'encore plus important c'est comment on va accéder au contenu ? Ce qui est important n'est sûrement pas la bouteille, c'est l'eau ou le vin mais ce qu'il y a de plus important encore c'est le tire – bouchon.

Et alors ce qui est intéressant c'est que dans le rapport Théry, il y a une petite phrase qui, à mon avis, n'est pas mise en évidence qui dit : « *parallèlement aux nouvelles techniques de dialogue, les développements réalisés dans le domaine des langages naturels, sous l'impulsion de France Télécom devront être suivis pour faciliter la recherche d'information dans les grandes bibliothèques accessibles en ligne* ».

Dans le rapport Thévy, il est effectivement reconnu que : « *la linguistique, l'analyse du langage naturel, seront une clé pour les autoroutes de l'information* ». Il faut un peu la chercher cette phrase, mais elle y est.

Je crois donc qu'on peut dire que le traitement automatique du langage transforme les données en informations, et ça pour moi c'est la clé. C'est à ça que sert le traitement automatique du langage, notamment du langage écrit qui est le type d'application sur lequel nous travaillons, et je voudrais vous montrer quelques exemples d'applications pratiques, opérationnelles pour la plupart, où la transformation de données en informations au sens où je l'entends est effective.

Qu'est ce que c'est qu'un système d'informations dans une entreprise ? C'est un processus ou une suite de processus qui vont permettre de créer des documents, de les ranger, pour les trouver ensuite, pour les réutiliser et pour les distribuer aux gens qui sont intéressés par ces documents, avec au – dessus un système qui gère et qui contrôle ce processus. Et ça c'est le schéma vrai de tout système d'information, y compris du système d'information humain. Quand on collabore avec ses collaborateurs dans une entreprise, on crée de l'information, on crée des documents, on les range, on les recherche, on les réutilise, on les distribue, et le support du système d'informations c'est le langage.

Donc le système d'informations informatisé doit comprendre le langage pour répondre à trois types de besoins :

- pour répondre aux questions que l'on a et pour y répondre en apportant des informations ;
- pour nous dire spontanément qu'est ce qui peut être une information pour nous, en scrutant des ensembles de données, en surveillant des ensembles de données, et dire : ici il y a une information qui nous intéresse ;
- pour aider à créer des documents, 80 % de l'information étant plus sous forme de documents que sous forme de données structurées.

Pour illustrer ça de manière concrète, je voudrais citer quelques exemples de projets et de réalisations concrètes.

Premier type de problème : répondre à des questions qui peuvent être posées soit par le grand public, soit par des professionnels dans la vie professionnelle. Nous avons réalisé par exemple avec France-Télécom une application sur l'annuaire électronique sur minitel qui est l'accès aux pages jaunes de l'annuaire électronique, le 11. Cette application est intéressante parce qu'elle a marché pendant plusieurs années avec des procédures traditionnelles sans linguistique et les taux de succès étaient très faibles, à peu près 40 % seulement des questions posées par les utilisateurs, les clients étaient incomprises. En ajoutant des procédures linguistiques, le système a été amélioré de manière très importante puisqu'aujourd'hui on a à peu près 95 % des questions qui sont bien comprises, et il y a 0,5 million de questions par jour qui sont analysées sur ce système. Donc on est vraiment là dans un cadre industriel : 500 000 questions en moyenne par jour. Le même type d'application peut être utilisé dans le cadre professionnel, par exemple, en ce moment nous réalisons une application avec l'institut national de la propriété industrielle pour permettre l'accès aux bases de données des brevets, directement par les PME qui, aujourd'hui, pour accéder aux brevets sont obligées de passer par des sociétés spécialisées en accès en information. Donc là, le langage naturel c'est la bonne forme pour trouver l'information.

Deuxième type d'application : le traitement automatique du langage permet de dire : *« Parmi une ensemble de données, qu'est ce qu'une information pour moi ? »* Et je donne là deux exemples, un exemple que nous faisons avec Aérospatiale, dans le domaine de la veille technologique où il s'agit d'analyser de grandes masses de documents, qui viennent soit de l'intérieur, soit qui sont téléchargés à partir de réseaux, et puis d'analyser ces documents, de les classer automatiquement en ensembles qui se ressemblent, pour permettre à des analystes de repérer des nouveaux rapprochements entre documents qui sont significatifs de l'émergence d'un nouveau besoin ou d'une nouvelle technologie dans leur champ de compétence, donc des problèmes, des applications liées à la veille technologique. Les applications d'un autre type de diffusion sélective d'information : nous avons fait une application avec EDF, où il s'agit encore là d'analyser des documents et de les envoyer aux ingénieurs qui sont concernés par le sujet du document. Donc chaque ingénieur s'est déclaré en disant : *« Voilà la liste des sujets qui m'intéressent »* et le système lui envoie les documents susceptibles de l'intéresser, avec un taux de succès qui n'est évidemment pas parfait, mais qui est meilleur que si on n'avait pas ce type de procédure. Et là, l'analyse du langage, c'est le bon moyen pour distribuer de l'information.

Enfin l'analyse du langage : les techniques de traitement automatique du langage naturel permettant d'aider à produire des documents, soit à rédiger – et je donne là deux exemples de système de vérification de langue contrôlée – vous savez que dans l'industrie lourde, notamment dans l'industrie aéronautique ou l'industrie automobile, il y a une tendance qui consiste à contrôler la forme des documents techniques, notamment des documents de

maintenance, essentiellement pour améliorer la lisibilité, s'assurer que les techniciens de maintenance qui vont réparer les avions n'ont pas plusieurs façons de comprendre les instructions, et d'autre part, afin de faciliter la traduction de ces documents qui sont énormes. Et donc là, le traitement automatique du langage apparaît comme étant la bonne forme pour créer.

Ce qui est très important – là j'ai un transparent qui est un peu confus, un peu fouillis (ALETH c'est le nom de notre boîte à outils) – c'est d'arriver à capitaliser les connaissances. Chaque application a coûté très cher, et un poste de coût important c'est le dictionnaire. Ce qui est donc important c'est d'arriver à faire en sorte que les investissements que l'on va faire sur le dictionnaire soient réutilisables d'une application à l'autre. Dans l'architecture des outils que l'on a mis au point, on a un ensemble de moteurs logiciels, de grammaires, et de dictionnaires qui sont indépendants des applications – c'est ça qui est important – et indépendants des modèles mis en œuvre dans les applications, et on a ensuite des systèmes de filtrage qui vont permettre de réutiliser les mêmes bases de données lexicographiques :

- dans des applications d'indexation et de recherche, – le premier type d'application que j'ai montré ;
- dans des applications d'aide à la traduction ;
- dans des applications d'aide à la rédaction ou de génération automatique de documents.

Et c'est le fait d'arriver à rendre le dictionnaire indépendant des modèles et des applications qui garantit la possibilité de capitaliser le développement. Ça va évidemment poser un problème parce que pour vendre les données lexicographiques indépendantes des applications, il faut un degré d'abstraction assez élevé puisque l'on va être obligé de donner beaucoup plus d'informations que le strict nécessaire pour une application, mais c'est le prix à payer pour la généralité.

Nous avons développé, notamment en partenariat avec un certain nombre de laboratoires industriels et de laboratoires de recherche universitaire, un dictionnaire électronique qui est indépendant des applications, comme je viens de le dire, afin de capitaliser les connaissances. Ce dictionnaire a été développé, notamment dans le cadre du projet Eureka Genelex, dont Stéphane Chadiron a dit un mot. Nous collaborons à l'initiative européenne *Eagles*, quand je dis « nous » je parle du consortium Genelex ; c'est-à-dire l'ensemble des participants de ce projet participe à un effort européen qui me paraît important, qui est connu sous le nom d'*Eagles*, qui est un effort visant à établir des recommandations, des préstandards, sur les formats d'échange de données lexicographiques, et ce que je voudrais souligner c'est que ceci représente un effort de plusieurs dizaines d'années/homme – en fait, ça frôle la centaine – donc ce sont des efforts importants qui ont déjà été faits, et on est, bien évidemment loin d'être arrivés au but. Je vous remercie !

Gestion électronique de documents en entreprises

Mon intervention se veut d'être d'un témoignage de quelqu'un qui a eu à gérer l'introduction d'un progiciel de gestion électronique de documents dans un environnement industriel.

Nous poserons les termes du problème, nous parlerons de ce que nous avons cherché à faire, des difficultés rencontrées, nous donnerons un aperçu du fonctionnement du progiciel et enfin nous terminerons sur les attentes des utilisateurs.

Laissez-moi tout d'abord vous présenter le cadre de ce projet. Il s'agit d'un bureau d'études pour l'automobile. Cela sous entend la quantité d'informations qui naît et s'échange depuis la phase de conception des véhicules jusqu'à leur sortie série.

C'est dans ce cadre qu'est né le projet « Accès aux savoirs techniques ».

À l'origine, nous avons cherché à savoir quelle relation existait entre une compétence technique et une recherche d'information. Auprès de nos ingénieurs on s'est rendu compte de l'existence de plusieurs démarches : la première est la plus courante est de faire appel à des connaissances acquises dans des projets déjà réalisés. Une autre démarche est d'avoir recours aux connaissances détenues par des experts que l'on appelle couramment dans le jargon du milieu « des pères techniques ». Ces « pères techniques » existent pour chaque métier pratiquement. Par exemple, il y a un père technique pour le collage... etc. La dernière démarche est d'avoir recours aux bases documentaires.

C'est de celle là, bien entendu dont je vais vous parler. Permettez-moi juste de rappeler la problématique. Toute personne à la recherche d'une information se trouve confrontée à un problème de délai d'obtention de l'information. Si la durée de l'obtention est supérieure à la durée de vie de l'information, il va s'en dire que c'est perdu d'avance. La disponibilité des experts sollicités pour répondre aux questions et le niveau de stocks de questions auxquelles ils doivent répondre constituent, également un élément important dans la fluidité qui peut se créer, dans une démarche de recherches d'informations.

La qualité de l'information va dépendre très souvent de la taille du réseau d'informations de la personne. Cette taille reste tributaire de son ancienneté dans le département, du poste qu'il occupe et surtout du nombre de personnes expertes susceptibles de connaître le sujet. Les moyens d'obtention vont du simple jeu de piste téléphonique où l'on connaît telle personne qui est en prise avec tel sujet et donc on va l'appeler pour obtenir une partie de la réponse que l'on escompte, jusqu'à la consultation depuis son bureau, via des bases de données documentaires, d'une liasse d'information structurée.

Le coût de la recherche est très variable. Si on utilise le téléphone, cela ne coûtera pas très cher. Dans le cas de la mise en place de données documentaires le coût reste élevé.

Qu'avons-nous chercher à faire ?

Élargir dans un premier temps le champs de recherche de nos ingénieurs. Cela a consisté à fournir à chaque ingénieur la liste d'autres ingénieurs qui travaillent dans le même champs d'action afin qu'il puisse consulter l'ensemble des travaux qui ont été faits dans le domaine qui le concerne. Nous avons cherché à voir comment l'on pouvait obtenir une information pertinente qui arriverait au bon moment. L'objectif est de se constituer un réseau formel d'information. Cela se traduira par la mise en place de bases documentaires installées sur un serveur et accessibles depuis n'importe quel poste client. L'expérience pilote d'un tel projet s'est déroulée dans deux secteurs assez demandeurs d'information : un secteur d'études de carrosserie et un secteur d'études de matériaux. Nous avons choisi un progiciel de gestion électronique de documents. Les documents à introduire dans la base concernaient soit des programmes véhicules, soit des définitions de produits, soit encore plus simplement des notes techniques tels des rapports d'essais ou autres.

L'architecture logicielle mise en place est la suivante : un serveur UNIX dans lequel sont stockés les documents, un réseau de clients PC connectés au serveur via un réseau. Le point d'achoppement d'un tel projet reste la phase, oh combien importante, de saisie de documents papier dans la base. C'est une activité sensible qui met en danger la réussite d'un projet de par son côté rédhibitoire ; le progiciel que nous avons choisi, met en œuvre des concepts basés sur une approche linguistique sophistiquée. L'existence de dictionnaires permet de mettre en œuvre le jargon du milieu. Intégrer le vocabulaire du métier évite d'avoir des « silences » lors d'interrogations des bases. Une normalisation des termes utilisés trouve là un bon champs d'application. Une approche statistique permet par ailleurs d'optimiser la recherche de documents pertinents. Cela se traduit par des classes de documents rangées par degré de pertinence. On peut, lors d'une interrogation, aller directement vers le document le plus pertinent. Par document pertinent, nous entendons le document le plus proche du point de vue sémantique des mots qu'on a introduits dans la question. Ce progiciel s'appelle Spirit. Quelles sont ses principales caractéristiques ? Les documents y sont structurés selon certains types de champs. Il y a trois modes d'accès à un document :

- soit par la référence du document. C'est le cas où on recherche un document dont on connaît le numéro de référence et donc on demande directement la référence de ce document.
- soit par le contenu structuré du document. On sait que tel document traite de tel sujet et donc dans ce cas l'interrogation portera sur le corps du sujet. C'est l'interrogation sur les champs d'un document.
- soit par le contenu libre (par opposition à structuré) du document. On interroge la base en texte libre.

Cet aspect est le plus novateur du progiciel. Je vais vous en parler un peu plus.

La boucle générale dans la question en langage libre est la suivante :

- dans un premier temps, vous avez une grille dans laquelle vous posez votre question en langage libre ;
- après validation, Spirit procède à une extraction des mots significatifs qui sont des mots informationnels tirés de votre question ;
- un processus interne normalise les mots, notamment mise à l'infinitif des verbes conjugués.... etc ;
- Spirit recherche les documents.

Pour vous témoigner de l'utilisation réelle de ce genre de logiciel en entreprise, je vais vous dire quels genres de problèmes, on rencontre dans leur mise en œuvre. Disons le tout de suite ce sont plus des problèmes liés à l'organisation qu'à des aspects techniques.

Il y a tout d'abord le problème de la qualité de la documentation papier. La reprise de l'existant se fait malheureusement avec des documents sous forme « papier ». Comme on n'a pas toujours le choix des documents à numériser, on est souvent confronté avec des documents dont la qualité n'est pas des meilleures. Leur passage au scanner couplés avec des logiciels de reconnaissance de caractères n'assure pas des taux de reconnaissance parfaits. Il faut savoir qu'un taux de reconnaissance de 90 % implique une non reconnaissance des 10 % ce qui est beaucoup dans ce genre d'exercice.

Le problème de l'hétérogénéité des plates-formes existe pour les grandes entreprises. Les ingénieurs travaillent généralement sur des postes de travail Unix, alors que les techniciens et autres managers sont généralement en prise avec des Macintosh ou des PC. Ces différentes plates-formes constituent une contrainte pour les éditeurs de logiciel. Les outils proposés doivent donc être en multi plates formes.

Le dernier problème et non des moindres et celui de la gestion de la confidentialité. Dans des secteurs sensibles, il faut pouvoir assurer la confidentialité aussi bien dans les différents secteurs concernés, qu'à l'intérieur même d'un secteur donné. Tout document doit littéralement savoir qui peut le consulter.

Ce sont là quelques retours d'information que nous transmettons à nos fournisseurs pour que dans les versions futures, nos besoins soient pris en compte.

Quelles sont nos attentes ?

Elles proviennent de différents points que nous avons abordés tout le long de cet exposé. Notamment :

- améliorer le traitement linguistique ;
- améliorer le traitement statistique – dixit les éléments du transparent précédent ;
- améliorer la reformulation.

Une dernière observation dans la mise en service de Spirit est la suivante : les techniciens sont des gens qui posent des questions plus directes, plus courtes, la plupart du temps sous la forme de mots-clés. Cela s'explique par le fait qu'ils ont une meilleure connaissance des documents qu'ils ont, au demeurant, la plupart du temps créés. Dans ce cas les questions sont mieux ciblées. Quand à ceux que l'on désigne communément sous le vocable de managers ou de décideurs, l'objectif de leur demande d'information reste d'avoir une idée globale. Leurs questions sont plus évasives, plus longues et généralement en langage libre. C'est dans ce cadre que les améliorations préconisées plus haut pourront aider les décideurs à obtenir également des réponses pertinentes.

J'espère que ce témoignage pourra intéresser ceux d'entre vous, désireux aujourd'hui de mettre en place un système de gestion électronique de documents dans le cadre d'une entreprise.

Du correcteur orthographique au correcteur grammatical intelligent

Ma présentation portera sur le passage du correcteur orthographique au correcteur intelligent. D'abord, un peu de publicité... La société Machina Sapiens est une société spécialisée dans le domaine de l'intelligence artificielle, particulièrement en traitement de la langue naturelle et dans les applications de la technologie de la programmation par objets.

Parmi nos premières réalisations dans le domaine du traitement des langues naturelles, nous avons produit plusieurs logiciels d'apprentissage du français dont Scarabée qui a été diffusé en France par Nathan sous le nom de Saga. Le didacticiel Saga était un système de jeux d'aventures avec lequel l'étudiant communiquait en tapant de courtes phrases en langue naturelle. C'était notre première expérience d'un analyseur de la langue naturelle.

Par la suite nous avons réalisé des systèmes multimédias pour l'enseignement du français comme langue seconde dont le système Elmo qui a été réalisé sur mesure pour le ministère de la Défense nationale.

Exploratexte qui a bénéficié d'une licence nationale du gouvernement français, est un tuteur intelligent pour l'apprentissage de la grammaire française. D'ailleurs ce produit a été adapté pour être utilisé par des clientèles de langues allemandes et anglaises pour l'apprentissage du français. Les menus présentés et des explications générées par le système ont été adaptés à l'allemand et à l'anglais. Nous travaillons à des adaptations pour l'espagnol et le japonais.

Le produit principal de Machina Sapiens, demeure toutefois le Correcteur 101, qui est un outil de correction du français reposant sur des technologies d'intelligence artificielle et de génie linguistique.

Mon exposé se concentre sur l'histoire afin de dégager les tendances de l'évolution des outils d'aide à la rédaction, et particulièrement de la correction grammaticale. Je peux affirmer que les outils de correction grammaticale sont devenus indispensables pour que

le français demeure une langue vivante. En effet, si on regarde ce qui existe du côté de l'offre en langue anglaise, il y a des outils linguistiques qui existent, des correcteurs qui marchent, alors qu'en français, jusqu'à tout récemment, les outils étaient plutôt de piètre qualité.

Au Québec, il y a même des gens qui hésitent, qui préfèrent, dans notre contexte nord-américain, écrire en anglais parce que c'est plus facile, plus simple, et qu'il y a des outils linguistiques performants comme des correcteurs. On dit que la nécessité crée l'outil. Par exemple, les Québécois ont inventé la motoneige, parce qu'ils avaient des hivers très rigoureux. Parce qu'ils sont plus sensibles aux problèmes reliés à la qualité de la langue française, les Québécois ont créé les premiers outils de correction grammaticale du français.

Donc l'histoire débute avec la première génération de correcteurs, les correcteurs orthographiques ou lexicaux, qui étaient dérivés de produits qui avaient eu du succès avec la langue anglaise. En général, les technologies du côté de la langue anglaise ont peu évolué puisque le besoin se faisait moins sentir. La langue anglaise étant moins complexe, on peut se contenter d'un simple correcteur orthographique ou lexical.

Ensuite, il y eut une deuxième génération de correcteurs qui a été inaugurée par un produit issu de nos compatriotes et compétiteurs de Logidisque, cette innovation s'appelait Hugo. On parle ici de correcteurs grammaticaux heuristiques ou à couverture locale. Il y avait également GrammR, un autre produit québécois réalisé cette fois par la société de John Chandioux. Enfin, nous arrivons à la troisième génération, celle des analyseurs grammaticaux, c'est-à-dire des systèmes capables d'une analyse détaillée de la langue naturelle pour faire des corrections. Le Correcteur 101 est premier produit de cette troisième génération.

Je reprends donc chaque élément rapidement :

Les correcteurs orthographiques ou correcteurs lexicaux sont basés sur une liste de mots, et l'algorithme d'analyse procède à une simple comparaison avec chaque mot de cette liste. Il n'y a donc aucune connaissance linguistique en tant que telle. Ce type de correcteur sert uniquement à la détection des fautes de frappe. On voit rapidement les limites de ce genre de système pour la langue française, car si la faute est dans la liste de mots du dictionnaire, on n'arrive pas à la détecter. Les performances, selon les textes, peuvent aller jusqu'à 40 % à 50 %. Pour aller plus loin, il va falloir développer une technologie plus proprement linguistique.

C'est là qu'est arrivée la deuxième génération, les correcteurs grammaticaux heuristiques. C'est-à-dire des correcteurs qui font des analyses locales ou partielles, puisque c'est assez difficile de faire l'analyse globale d'une phrase quelconque. Ce sont des systèmes heuristiques capables d'identifier les fautes les plus fréquentes grâce à des analyses partielles ou locales, souvent réalisées par des algorithmes apparentés aux

automates. Ces analyseurs permettent de faire des corrections de segment dans le genre, « les petit chiens », c'est-à-dire suffisants pour contrôler les accords locaux. Cette innovation introduite par les correcteurs de seconde génération peut être améliorée jusqu'à un certain point pour tenter d'analyser complètement les phrases.

Ce qui fait l'objet de la troisième génération de correcteurs, ce sont les analyseurs grammaticaux capables de faire l'analyse complète et précise de phrases quelconques. Ces analyseurs reposent sur des connaissances linguistiques approfondies. On parle ici de système à base de connaissances, ou de systèmes experts en grammaire française. Personnellement je préfère le terme système à base de connaissances, parce que l'expertise se trouve dans les précis de grammaire. Évidemment, il y a toute une partie de ces connaissances qui relèvent du traitement de la langue naturelle, qui a aussi ses propres règles.

Voilà pour un bref historique des logiciels correcteurs pour le français.

Parmi les autres applications du traitement de la langue naturelle, je passe rapidement puisqu'on en a parlé dans d'autres séminaires, je souligne l'interrogation de banques de données en langue naturelle qui présente un intérêt particulier en permettant un accès direct à l'information.

Le repérage de documents, la génération automatique de textes, l'extraction de notions, la simplification de textes, la génération de résumés, les outils d'aide à la traduction, la postsynchronisation et la reconnaissance vocale sont d'autres applications à venir. Ces applications ne sont pas listées en ordre de complexité, mais en ordre de réalisation potentielle.

Globalement, on observe un passage du traitement de textes au traitement de documents et d'une culture, je dirais, littéraire, à une culture audiovisuelle. C'est ce qui explique que le français n'est pas la seule langue menacée. Partout dans le monde, les gens de la génération des moins de trente ans éprouvent les mêmes problèmes de maîtrise de la grammaire et la syntaxe de leur langue maternelle. Je pense que ce phénomène traduit le passage d'une culture de l'écrit à une culture audiovisuelle. Ce sont toutes les langues du monde qui s'appauvrissent.

Au niveau technologique, on passe du traitement de textes (ou du traitement de chaînes de caractères) au traitement de l'information, et même au traitement des connaissances. Et là on touche à la sémantique et à l'intelligence artificielle avec un grand A.

On assiste aussi à l'intégration de modules intelligents et à l'apparition de systèmes assistés, plutôt que des systèmes entièrement automatiques. Les gens qui voulaient faire des systèmes entièrement automatiques ont un peu déchanté. Ce qui est clair, c'est qu'on ne trouve plus un représentant commercial à l'heure actuelle pour oser en parler...

De plus, on parle essentiellement de systèmes qui vont assister l'humain et qui n'ont pas de connaissances du sens commun. Par exemple, si je dis « le panier de pommes rouges »,

est-ce le panier qui est rouge ou les pommes qui sont rouges ? Évidemment, c'est la personne qui a écrit le texte qui le sait. On ne va pas programmer une énorme base de connaissances sur le sens commun pour trancher un problème dont seul le scripteur connaît la réponse. Voilà pourquoi on se dirige vers la correction assistée et la traduction assistée où le sens commun est fourni par l'utilisateur humain. C'est une étape que l'on peut alléger, car le système « intelligent » peut proposer les différentes possibilités. Mais c'est l'utilisateur humain qui, en bout de ligne, prendra la décision finale.

Nous verrons de plus en plus d'intelligence artificielle dans les outils bureautiques, et elle sera de plus en plus intégrée aux systèmes existants. Par exemple, les logiciels de traitement de texte bénéficieront de correcteurs et d'outils d'aide à la rédaction. Ce seront davantage des systèmes assistants que des systèmes entièrement automatiques.

Je conclurai par un petit aparté sur l'inforoute. D'ailleurs j'aime bien le terme d'inforoute qui traduit mieux la réalité que le terme d'autoroute électronique qui réfère à des fils ou à de la quincaillerie.

Machina Sapiens est actuellement impliqué dans un projet relié à un problème important sur l'inforoute Internet. Le problème, nous l'avons baptisé « infobésité », c'est-à-dire la surcharge, le trop-plein d'information. Tous ceux qui sont abonnés aux listes de nouvelles d'Internet savent bien ce que je veux dire. De retour à Montréal, je vais probablement avoir reçu quelques centaines de messages des différents groupes d'intérêt auxquels je suis abonné. Heureusement, InfoScan va m'aider à dépouiller tout cela.

Le remède à l'infobésité, c'est InfoScan, un logiciel qui filtre et classe les documents, particulièrement le courrier électronique, à partir de profils d'intérêt exprimés par des mots-clés. InfoScan, c'est un assistant intelligent qui filtre le courrier électronique reçu d'Internet. L'idée d'InfoScan est de se mettre au régime en éliminant le surplus d'information.

L'utilisateur n'a même pas à trouver les mots-clés dans ses documents. InfoScan se charge de les extraire dans des textes désignés par l'utilisateur. À l'avenir, InfoScan va repérer des textes qui se rapprochent de ces textes choisis.

InfoScan présente deux innovations importantes : l'extraction automatique des mots – clés en trois langues (anglais, espagnol et français) et la visualisation des documents – filtrés – à partir de ces mots-clés. Nous avons donc développé un concept d'interface pour visualiser d'un coup d'œil les documents les plus pertinents.

Plus Internet progresse dans les entreprises, plus le besoin d'outils comme InfoScan va se faire sentir. Il y a des masses énormes d'information dans le réseau Internet, les utilisateurs d'Internet passent des heures innombrables à s'échanger des messages et à lire des messages rarement pertinents. Bref, les internautes ne travaillent plus ! On a donc un urgent besoin d'outils pour contrer cette infobésité. Les applications potentielles d'InfoScan sont le filtrage des nouvelles d'Internet, le repérage d'informations, la sélection

même de films ou de documents multimédias. On touche là au problème classique des banques de données d'images et de films. Il semble bien que ces systèmes de repérage d'images vont utiliser l'indexation de textes descriptifs.

À court terme, on parle essentiellement de filtrer et de classer le courrier électronique reçu d'Internet.

