ÉTUDE SUR LE TRAITEMENT INFORMATIQUE DU FRANÇAIS ET DE SES LANGUES PARTENAIRES

Services et technologies d'ingénierie linguistique d'Alis

Avant-propos

Cette étude est le fruit d'un protocole d'accord entre l'Agence de la Francophonie et Alis Technologies inc. dans le cadre des programmes de l'Agence « La Francophonie dans le monde – OTAF ».

Le mandat donné à Alis Technologies par l'Agence est le suivant : « *réaliser une étude sur le traitement informatique du français et des langues partenaires de la Francophonie pour la traduction dans le cadre des nouvelles technologies : état actuel de l'art, évolution technologique prévisible, contraintes et possibilités, quelles actions pour la Francophonie, quelles priorités…?* »

Le présent rapport est le résultat de la réalisation de ce mandat.

alis TECHNOLOGIES

NOVEMBRE 1997

Table des matières

1. Introduction

Un sondage paru récemment 4 nous apprenait que le français occupait la quatrième place dans les contenus inforoutiers. Place somme toute honorable. Aujourd'hui, on trouve également un grand nombre d'outils informatiques destinés au grand public francophone et qui traitent convenablement les textes en français. Le marché offre une kyrielle de correcteurs grammaticaux de plus en plus perfectionnés. Microsoft sort les versions de la plupart de ses logiciels en français en même temps que ses versions anglaises, ceux-ci viennent équipés d'outils linguistiques adéquats (correcteur, thésaurus, condenseurs de texte[1], etc.).

Faut-il pour autant s'assoupir ou, à l'inverse, céder aux plaintes constantes de ces Cassandre qui nous ressassent quelque antique plainte et annoncent sans cesse la marginalisation prochaine du français ?

Pour répondre à ces questions, nous avons voulu, dans la première partie de cette étude, effectuer un bref survol de l'état des lieux. On y découvrira que si un faisceau de facteurs favorables, tels que les forces du marché[2] en présence, la conscience linguistique des consommateurs et des producteurs, la proximité graphique du français par rapport à l'anglais et même certaines lois, a permis au français de « maintenir son rang » dans le traitement linguistique, d'autres facteurs techniques prépondérants font craindre qu'il y ait péril en la demeure.

En effet, il faut bien admettre que l'établissement des normes de l'inforoute[3] par un cénacle anglo-saxon, que cette genèse nord-américaine[4] en quelque sorte n'assure toujours pas la réception non corrompue d'un message écrit en français. Au demeurant, il est toujours impossible aujourd'hui sur ces inforoutes d'écrire le nom de son correspondant en français ou de publier à l'intention de consommateurs francophones une adresse naturelle pour eux et qui enfreindrait les règles américaines en usant de diacritiques. Il est évident que cette limitation est d'autant plus débilitante pour les membres de la Francophonie que leur langue connaît des écarts graphiques plus grand encore par rapport à l'ASCII.

Ce ne sont là que les signes précurseurs les plus visibles de maux qui nous guettent. Nous assistons actuellement à une formidable accélération dans le domaine des technologies de l'information. Accélération tant au niveau des échanges, de la taille des réseaux que du volume des données. Cette accélération, facteur de sélection 4, et la pléthore d'information qui l'accompagnera nécessiteront de plus en plus d'outils intelligents capables d'aider l'humain à débusquer l'information pertinente dans un dédale de données. Et qui dit intelligent dit, en fin de compte, capable de « comprendre » les textes stockés et les requêtes de l'homme.

---

[1] Cf. la section 2.4.1, Résumeur automatique, pour la distinction entre un condenseur de texte et un résumeur.

[2] La France et le Canada sont tous deux des forces économiques mondiales.

[3] Toutes les remarques concernant l'Internet s'appliquent également aux réseaux d'entreprises, les intranets qui réutiliseront massivement toutes les technologies développées pour l'Internet.

[4] « Au début était le Verbe et le Verbe était en ASCII majuscule. »

On peut penser que les outils destinés au grand public seront fournis en français par les éditeurs américains. L'expérience nous apprend cependant que les forces du marché ne sauront nous prémunir d'outils « standards » mais incapables d'assurer un traitement correct du français. En outre, on peut légitimement penser que, sans un investissement supplémentaire, certains de ces outils en français seront moins puissants, moins polyvalents, moins productifs ou encore ne feront que renforcer les forces centripètes du tout vers l'anglais. La Francophonie, tant d'expression maternelle que de culture, ne pourrait que pâtir de ces obstacles.

Comment conjurer ces menaces ? C'est l'objet de la dernière partie de notre étude. Nous y proposons une série de recommandations que nous pensons réalistes et gérables. Certaines même, dans le domaine de la normalisation, peuvent être mises en œuvre avec de modestes budgets d'autant plus que l'on peut compter sur des alliances naturelles : la Francophonie n'est pas la seule à être affligée de certains des désavantages énumérés. Enfin, nous avons également présenté des pistes de solution qui permettraient d'impliquer les pays du Sud dans ce domaine complexe, premiers pas opérationnels et pratiques vers une prise en compte de leurs langues, mesures qui prennent pour appui le français.

2. État des lieux

2.1. Les bases du traitement informatique

1.1.1. Le codage des caractères

La base de tout traitement informatique du langage écrit est le codage des caractères de ce langage. En toute généralité, il est parfois nécessaire de préciser *quels sont* vraiment les caractères de l'écriture, une question qui peut sembler étrange aux utilisateurs d'alphabets simples comme l'alphabet latin, mais qui se pose sérieusement dans le cas d'autres systèmes d'écriture lorsqu'il s'agit d'en créer un codage informatique utile. Pour prendre un exemple parmi les langues partenaires du français, considérons l'écriture arabe : dans cette écriture, chaque lettre peut prendre des formes différentes selon la position dans le mot (initiale, médiale, finale ou isolée) ; la question est alors de savoir s'il faut coder chacune de ces formes séparément, ou seulement coder l'identité de chaque lettre et laisser aux logiciels d'affichage le soin de choisir la bonne forme en fonction du contexte.

Le problème du codage s'est posé très tôt dans l'histoire de l'informatique, mais les solutions pour diverses langues ne sont pas apparues en même temps. La première langue codée, bien qu'imparfaitement, a été l'anglais – pour des raisons historiques évidentes – avec des codes à 5, 6 et 7 bits. Un de ces codes, l'ASCII à 7 bits, s'est imposé comme un standard quasi universel, les seules exceptions étant les codes EBCDIC utilisés sur les grosses machines d'IBM.

L'ASCII ne permet de coder que 128 caractères, dont un certain nombre de caractères de commande et plusieurs signes de ponctuation et symboles divers, ce qui ne laisse pour la langue comme telle que les 26 lettres de l'alphabet latin de base, en haut et bas de casse. De nombreux codes ont été développés pour coder d'autres caractères pour les langues ne pouvant se contenter du latin de base, mais la constante est qu'ils reprennent presque tous l'ASCII comme base, ce dernier restant donc incontournable et universel. L'ASCII a d'ailleurs été rapidement adopté comme base de l'Internet, le réseau des réseaux que l'on ne peut plus ignorer aujourd'hui quand on parle de traitement informatique.

Pour le français, de multiples codages ont été proposés et implantés au cours des ans. Les premières solutions devaient tenir compte de la limitation à 7 bits de beaucoup de matériels et de logiciels ; on a donc vu des codes à 7 bits basés sur ASCII, mais remplaçant quelques symboles de moindre utilité par quelques-uns des caractères accentués nécessaires au français. Cette manière de faire, fort insuffisante, a été standardisée dans la norme ISO 646.

Sont ensuite apparus les codes à 8 bits, tels que les diverses pages de code IBM, Apple ou Microsoft et la série de normes ISO 8859. Ces codes sont généralement adéquats pour le français – du moins ceux qui sont conçus pour cette langue – mais le problème est qu'il y en a trop ! Diverses machines utilisent des codes différents, généralement sans étiquetage adéquat (il faut deviner quel code est utilisé dans chaque cas particulier), ce qui complique singulièrement l'échange d'information entre utilisateurs francophones, que ce soit au

moyen de vulgaires disquettes, par un disque partagé sur un réseau ou via l'Internet. Sur ce dernier, le code ISO 8859-1 (mieux connu sous le nom d'ISO Latin-1) est devenu un standard *de facto* pour les utilisateurs francophones[1], justement pour pallier ce problème de manque d'étiquetage ; la messagerie, les forums de discussion, les pages Web utilisent ce code, à charge pour les machines utilisant d'autres codes en interne de s'occuper du transcodage à l'émission ou à la réception.

L'ISO Latin-1 a un petit problème pour le français : un incident malencontreux dans le processus de normalisation y a causé l'absence des caractères **œ**, **Œ** et **Ÿ** nécessaires à une orthographe vraiment correcte[2]. On retrouvera donc *coeur* au lieu de *cœur*, et il n'est pas possible d'écrire *L'Haÿ-les-Roses* en majuscules en Latin-1. Ce problème peut sembler mineur – tout lecteur reconnaîtra *coeur* sans difficulté – mais il peut avoir des conséquences réelles : par exemple, une recherche sur le Web du mot *cœur* correctement orthographié ne trouvera rien. Ce problème est en voie d'être réglé, la solution étant ironiquement un effet secondaire de l'introduction de la nouvelle monnaie européenne, l'euro, ou plus précisément de son symbole. Sous la pression de la Commission Européenne, une nouvelle partie de l'ISO 8859 (partie 15, Latin-0) est en cours de normalisation, et on a profité de l'occasion pour y ajouter les caractères manquant pour le français (et le finnois) en plus du symbole de l'euro. La grande question est de savoir jusqu'à quel point ce Latin-0 pourra prendre la place privilégiée aujourd'hui occupée par le Latin-1, et comment les problèmes de conversion de données seront réglés.

La situation du codage des langues partenaires du français est inégale. Les langues d'Europe, surtout occidentales, n'ont pas de réels problèmes, hormis ceux d'étiquetage et de conversion qui affligent aussi le français. Le roumain est moins bien servi, le code ISO 8859-2 censé le couvrir ayant été, à l'instar du Latin-1 pour le français, victime d'une erreur de conception. Les codages des langues africaines, du vietnamien et du khmer sont nettement plus problématiques ; en fait, dans le cas du khmer, il n'existe aucun codage standard ni aucun qui soit techniquement satisfaisant.

Dans un avenir proche, la norme de codage ISO 10646 4 (équivalente au standard industriel Unicode) prendra la relève des codes à 7 et 8 bits dans les systèmes modernes. Ce code est destiné à reprendre les caractères de toutes les écritures du monde en un seul codage uniforme, les plaçant toutes sur un pied d'égalité et annihilant de ce fait les éternels problèmes d'étiquetage et de conversion de jeux de caractères. Ce nirvana ne sera bien sûr atteint qu'après une longue période de transition, mais des systèmes Unicode existent déjà, et il est agréable de constater que le français et presque toutes ses langues partenaires y sont parfaitement à l'aise.

1.1.2. La saisie (claviers)

Il y a peu à dire sur les claviers permettant la saisie du français, si ce n'est qu'il en existe une assez grande variété. Les différences sont issues de traditions nationales, mais aussi

[1] Il en est de même pour les utilisateurs d'une poignée d'autres langues européennes bien servie par l'ISO Latin-1 : espagnol, italien, allemand, danois, etc.

[2] Rappelons qu'en français des mots comme *cœur* s'écrivent avec l'e dans l'o, qui ne représente qu'un seul son, alors que d'autres comme *coexister* ne lient pas l'e et l'o. Il est clair en rétrospective que les normalisateurs de l'ISO Latin-1 n'étaient pas au courant de ce fait.

de différences gratuites entre constructeurs qui nuisent à l'utilisateur passant d'un clavier à un autre. La norme ISO 9995 4 cherche à mettre un peu d'ordre dans ce décor, en imposant des règles de conception de claviers qui assurent une certaine uniformité, mais elle ne précise pas tout et n'est pas encore très suivie.

L'introduction de l'euro aura encore ici un effet, puisque que les claviers européens devront pouvoir saisir le symbole euro ; on peut espérer que les fabricants profiteront de l'occasion pour se rapprocher de l'ISO 9995 et ainsi améliorer l'uniformité à l'intérieur des familles (AZERTY, etc.) de claviers.

1.1.3. Le tri et le classement

Le tri étant un des problèmes les plus anciens et rebattus de l'informatique, il est quelque peu surprenant que le tri correct du français n'ait été correctement résolu qu'assez récemment. Pendant les premières décennies de l'informatique, la plupart des efforts ont porté sur la mise au point d'algorithmes de tri efficaces, robustes et utilisables sur des machines à mémoire très limitée, tous prenant pour acquis la disponibilité d'une fonction de comparaison de chaînes de texte qui rend l'ordre relatif de deux chaînes. Mais cette fonction n'est pas élémentaire, puisqu'elle doit tenir compte de la casse (majuscules, minuscules), des accents et des caractères « spéciaux » (ponctuation, etc.) en respectant les règles de tri propres à chaque langue, et même à chaque application[1], en la matière. Ce n'est que vers la fin des années 1980 que le problème de la comparaison a été réglé, menant à la publication d'une norme canadienne 4 permettant le tri correct du français (et en prime de cinq autres langues européennes).

Un projet de norme internationale en matière de classement et de comparaison (ISO/CÉI 14651 4) est présentement en cours. Les deux objectifs poursuivis par les promoteurs du projet sont l'harmonisation des méthodes de classement de chaînes de caractères utilisées par les communautés linguistiques recourant à un même système d'écriture, de même que l'intégration de la norme de codage JUC[2] dans la future règle.

2.2. Les réseaux, leurs protocoles et l'impact sur le traitement du français et de ses langues partenaires

Nous mettrons dans cette section l'accent sur Internet, pour la bonne raison que ce réseau des réseaux domine aujourd'hui – et dominera encore demain – tout le développement technologique et toute la pratique courante des réseaux.. Nous pourrions discuter brièvement des réseaux locaux mais, s'agissant principalement de partage de fichiers et hormis quelques problèmes de « nommage » aussi présents sur Internet, il y aurait peu à dire et nous voulons garder cet état des lieux bref.

[1] Par exemple, on ne trie pas un annuaire téléphonique de la même manière qu'un dictionnaire, bien que l'ordre alphabétique de base soit évidemment le même.

[2] Le JUC est le Jeu Universel de Caractères de l'ISO/CÉI 10646 4, également connue sous le nom commercial d'Unicode.

On peut classer les protocoles d'Internet en trois grandes catégories :

- les protocoles de base, qui assurent le fonctionnement du réseau lui-même ;
- la messagerie, à savoir le courrier électronique et les forums de nouvelles[1] ;
- la recherche et la récupération d'information, ou, plus succinctement, les services d'information Internet.

Parmi les protocoles de base, on peut mentionner les suivants : `IP`, `TCP`, et `DNS`[2]. La messagerie repose sur les protocoles `SMTP` son extension et les forums de nouvelles sur `NNTP`. Du côté des services d'information, les principaux sont `FTP`, `Telnet`, et `HTTP` qui est le protocole de base du W3. Chacun de ces services fait l'objet d'au moins un protocole standard de l'Internet, et se distingue par ce protocole, ainsi que par son intéractivité et par le type de données qu'il permet de traiter.

Du point de vue du soutien des langues, notamment du français, il n'y a pas grand-chose à dire sur les protocoles qui constituent vraiment la base, l'ossature de l'Internet : `IP`, `TCP`, `UDP` et `ICMP` et les divers protocoles d'aiguillage. *Grosso modo*, ils réalisent la fonction de base du réseau, à savoir permettre le transport d'information numérique d'un point de départ à une destination. L'aspect le plus important est qu'à ce niveau, on a prévu la *transparence*, c'est-à-dire que le réseau peut transporter l'information sans la modifier de quelque façon que ce soit, sous forme d'une séquence d'octets de 8 bits. `IP` associe à chaque machine un numéro tel que *199.84.165.125* qui sert d'adresse unique pour joindre cette machine. `TCP` permet d'établir des connexions entre machines (c.-à-d. chaque machine sait qu'elle est connectée à l'autre, un peu comme lors d'un appel téléphonique).

Les problèmes commencent dès qu'apparaît la notion de texte, c'est à dire dès qu'on veut associer une signification autre que numérique aux octets transmis. À un niveau assez fondamental, considérons le `DNS`, qui sous-tend la plupart sinon la totalité des autres services de l'Internet. La raison d'être de ce système est de permettre l'utilisation de mots à la place des numéros que le protocole `IP` associe à chaque machine : *ampère.alis.ca* est nettement plus sympathique que *199.84.165.125*. Mais il y a un os : le e accent grave *d'ampère* est interdit, ne faisant pas partie du pauvre répertoire ASCII. C'est donc dire que les seuls anglophones peuvent jouir du plein bénéfice mnémonique de ce système. Un francophone devra se contenter d'*ampere* pour honorer l'homme dont les découvertes en électricité ont à terme mené au développement d'Internet [ou encore choisir de rendre hommage à Chappe, créateur du premier réseau de communication numérique (1794)]. Un sinophone sera limité à un compromis encore plus défavorable.

Une autre manifestation de l'hégémonie de l'ASCII apparaît dans les noms de fichiers informatiques, souvent transmis sur l'Internet pour, par exemple, demander la transmission d'un de ces fichiers. Beaucoup de systèmes d'exploitation ne permettent pas l'utilisation de caractères autres que ceux de l'ASCII dans les noms de fichier ; ceci ne concerne pas directement l'Internet, mais s'impose quand même aux utilisateurs. D'autre

[1] aussi connus sous divers vocables comme *Usenet*, les *inforums*, les *news*, les *groupes de news*, les *niouzes*, etc.

[2] *Domain Name System*. C'est un système distribué et hiérarchique qui permet d'associer un nom, plutôt qu'un simple numéro sans valeur mnémonique, à toutes les machines reliées à l'Internet.

part, même si certains systèmes permettent l'emploi de noms plus riches, de tels noms sont pratiquement interdits de séjour sur l'Internet pour la raison suivante : les logiciels devant traiter ces noms présupposent l'ASCII, et il n'existe pas dans les protocoles Internet de moyens de faire savoir à l'autre bout quel codage autre que l'ASCII a été utilisé. *Ergo*, pas question d'intituler *Ampère.txt* un panégyrique que l'on voudrait largement accessible sur l'Internet.

La même situation se répète dans les URL[1], qui sont les adresses universelles de ressources sur Internet. Un URL se compose généralement d'un nom de machine, soumis aux restrictions du DNS mentionnées ci-dessus, auquel est accolée une chaîne de caractères représentant le nom de la ressource sur la machine en question. Mais les standards Internet régissant la construction des URL ne tiennent évidemment compte que de l'ASCII, ce qui interdit encore une fois d'utiliser des noms vraiment significatifs en toute autre langue que l'anglais. Qui plus est, les URL sont aussi utilisés – il s'agit d'un artifice technique – pour la soumission du contenu d'un formulaire HTML à un serveur Web. La limitation à l'ASCII s'applique encore en principe, mais ici la pratique dépasse de loin la norme, puisque tous les jours des internautes écrivent leur nom, celui de leur ville ou un mot-clé pour une recherche dans des formulaires HTML, que ce soit en français, en arabe ou en japonais. Le problème en est un ici de fiabilité, l'insuffisance des standards ne permettant au mécanisme de fonctionner hors ASCII que par pur hasard[2]. La standardisation des URL non-ASCII est en cours, mais elle se heurte à une très grande inertie, et il n'est pas exclus que des pressions de nature politique, dépassant les cercles techniques, ne soient nécessaires pour arriver au résultat.

Le courrier électronique Internet est fondé sur le protocole `SMTP` et sur un format de message standardisé connu sous le nom du document constitutif (*format RFC 822*[3]). Ces deux standards Internet spécifient explicitement l'usage exclusif du jeu de caractères ASCII, ce qui ne permet strictement de transmettre que du texte simple en anglais. Toute autre transmission exige un codage particulier camouflant le message sous la forme de courtes lignes de caractères ASCII.

Une extension (`ESMTP` 4) permet sous certaines conditions la transmission de caractères à 8 bits, donc autre qu'ASCII ; malheureusement, cette méthode ne fonctionne pas encore vraiment en pratique, pour plusieurs raisons. D'une part, la transmission à 8 bits reste optionnelle, et n'est donc pas universelle. D'autre part, elle n'est permise qu'après négociation fructueuse entre les deux serveurs de messagerie impliqués, mais il est déjà trop tard, au moment où cette négociation peut se tenir, pour que l'agent-utilisateur de messagerie puisse décider du codage à utiliser ; l'expéditeur reste donc forcé de surcoder tout ce qui n'est pas ASCII, pour pallier le cas (encore fréquent) d'une négociation infructueuse, et `ESMTP` ne règle rien en pratique[4].

---

[1] *Uniform resource locator.*

[2] Le mécanisme fonctionne si client (c-à-d fureteur) et serveur font la même hypothèse quant au jeu de caractères utilisé. Hypothèse souvent vérifiée lorsqu'une seule langue est en jeu, mais qui s'effondre dans un grand nombre de cas.

[3] *RFC* signifie *Request For Comment*. Les RFC constituent une série de documents qui contient, entre autres, tous les standards de l'Internet.

[4] En fait, le serveur expéditeur pourrait lui-même surcoder en cas de négociation infructueuse avec le serveur destinataire, et même doit le faire d'après le RFC idoine, mais le manque d'universalité de serveurs capable de ce faire, ainsi que certains autres problèmes techniques, justifient notre affirmation.

Une autre extension, MIME (4 4 4 4 4), est nettement plus porteuse. Elle spécifie des méthodes de codage normalisées, mais surtout formalise la transmission d'information sur le codage effectivement utilisé à l'expédition, permettant au destinataire de décoder le message sans incertitude. MIME permet en principe de transmettre du texte en n'importe quel jeu de caractères, et donc en n'importe quelle langue, en plus de permettre la transmission fiable d'images, de son, de vidéo, de fichiers binaires d'applications, etc. Le problème de MIME a été au début son manque d'universalité : tous n'avaient pas les logiciels appropriés. Ce problème, heureusement, est en bonne voie de disparaître : MIME est répandu au point que celui qui ne l'a pas se sentira coupable, plutôt que celui qui a transmis un message exigeant MIME. Il reste toutefois que nombre des logiciels MIME en circulation ne sont pas complets ou ont des défauts qui affectent le français[1], les éditeurs portant plus d'attention aux aspects multimédia de MIME qu'aux soutien des langues.

La situation est un peu plus rose du côté des forums, mais seulement pour certaines raisons historiques que l'on pourrait qualifier d'accidentelles. En effet, les normes régissant cette application sont tout aussi restrictives que dans le cas du courrier, mais sont appliquées avec moins de rigueur. Le format d'un message Usenet est le même que celui d'un message de courrier (RFC 822), à quelques écarts près ; officiellement, seul l'ASCII est admissible. Le transport est assuré sur l'Internet par le protocole NNTP, qui lui aussi spécifie l'ASCII exclusivement. Heureusement, cette stipulation est ignorée par les serveurs NNTP les plus courants, qui laissent tranquillement passer les caractères à 8 bits, ce qui permet en pratique la propagation de messages autres qu'ASCII. Il n'en reste pas moins que la survivance de serveurs imposants la norme à 7 bits, ainsi que le manque d'étiquetage du jeu de caractères utilisé, rend cette pratique moins qu'idéale et universelle.

2.3. Recherche et filtrage d'information

Avec l'avènement de l'Internet, le problème du repérage d'information, domaine réservé naguère encore à une caste spécialisée de bibliothécaires ou de documentalistes, est devenu une réalité omniprésente pour tout internaute. En effet, qui n'a ressenti une énorme frustration en contemplant les résultats fournis par une recherche effectuée sur la Toile basée sur l'emploi de mots-clés ? En théorie, cependant, il est dorénavant possible de mener une recherche intelligente, à partir de requêtes formulées en langage naturel (en français courant par exemple). Grâce à ce type de techniques on peut donc désormais repérer des concepts plutôt que celui de simples chaînes de caractères.

La recherche et le filtrage d'information constituent deux variantes d'une démarche qui consiste à extraire d'une masse (peut-être très grande) de documents un sous-ensemble jugé pertinent relativement à un besoin exprimé.

Le terme « recherche d'information » (ou « recherche documentaire ») concerne une situation dans laquelle la collection de documents est plus ou moins fixe alors que le besoin des utilisateurs varie. Les utilisateurs formulent des « requêtes » et le système doit

[1] Par exemple les caractères accentués dans les en-têtes – champs objet, destinataire, etc. – posent souvent problème, étant l'objet d'un sous-protocole délicat, complexe et sensible aux réalisations imparfaites.

chaque fois extraire de la collection complète le sous-ensemble des documents qui sont pertinents relativement à la requête.

Par contraste, le terme « filtrage d'information » réfère plutôt à une situation dans laquelle le besoin informationnel (la requête) est fixe, alors que la collection statique de documents est remplacée par une source continue. Par exemple, un consommateur pourrait vouloir obtenir, jour après jour, tous les documents relatifs à l'industrie nucléaire qui sont transmis sur un canal d'information précis.

Qu'il s'agisse de recherche ou de filtrage, les techniques utilisées seront en général les mêmes. La requête sera exprimée soit: 1) comme un ensemble de mots clés, peut-être organisés par des opérateur booléens; ou 2) comme un bref énoncé en langue naturelle qui décrit le sujet d'intérêt; ou 3) comme un texte que le système doit prendre comme un exemple des textes pertinents. La tâche du système est toujours de sélectionner un sous-ensemble pertinent parmi l'ensemble des documents disponibles.

Aux fins de la recherche dans une collection plus ou moins fixe de documents, la collection fera l'objet d'un traitement préparatoire en vue rendre la recherche raisonnablement rapide. Habituellement, ce prétraitement consiste à représenter chaque document D par un vecteur de poids qui correspond au contenu lexical du document :

D_1 : { (mot_1, $poids_{1,1}$), (mot_2, $poids_{2,1}$), ... (mot_n, $poids_{n,1}$) }
D_2 : { (mot_1, $poids_{1,2}$), (mot_2, $poids_{2,2}$), ... (mot_n, $poids_{n,2}$) }

Typiquement, la requête sera, elle aussi, transformée en un vecteur de poids lexicaux :

R : { (mot_1, $poids_{1,r}$), (mot_2, $poids_{2,r}$), ... (mot_n, $poids_{n,r}$) }

La sélection des documents pertinents à la requête se fait alors en appliquant une mesure qui quantifie la similarité entre $\boldsymbol{R}$ et chaque document $\boldsymbol{D}_i$, et en retenant les documents pour lesquels la similarité mesurée dépasse un seuil préétabli.

La tâche de filtrage fonctionne de façon analogue. On associe un vecteur de poids à chaque nouveau document qui apparaît sur le canal, et on compare ce vecteur à celui ou ceux qui décrivent le profil d'intérêt du consommateur d'information.

Pour évaluer les systèmes de recherche et de filtrage d'information, on doit normalement disposer d'un petit nombre de requêtes typiques pour lesquelles les réponses correctes ont été fournies (par exemple par des documentalistes humains). La réponse correcte $\boldsymbol{C(R)}$ à une requête R est par définition un sous-ensemble précis de la base documentaire complète[1]. On soumet la requête R au système évalué et on compare la réponse du système $S(R)$ à la réponse correcte $C(R)$.

La comparaison fait normalement intervenir deux mesures différentes. La première de ces mesures est le taux de *rappel*. Si $|X|$ dénote la cardinalité de l'ensemble X, ce taux $\boldsymbol{Ra(R)}$ se définit comme suit :

[1] Dans le cas d'une tâche de filtrage, la base documentaire pertinente sera constituée par l'ensemble de tous les documents produits dans un certain intervalle temporel par la source à filtrer.

$$Ra = \frac{|S(R) \cap C(R)|}{|C(R)|}$$

En d'autres termes, on se demande dans quelle mesure les documents qu'il fallait choisir ont effectivement été choisis. Le taux rappel ***Ra(R)*** a pour dual le taux de *silence Si(R)* : par définition *Si(R) = 1 - **Ra(R)***.

À elle seule, cette mesure est peu informative puisqu'on peut obtenir un rappel parfait simplement en associant à ***R*** la base documentaire complète. Le taux de rappel prend plutôt son sens lorsqu'on le conjoint à la seconde mesure: le taux de *précision* ***Pr(R)***, qui se définit comme suit :

$$\mathrm{Pr} = \frac{|S(R) \cap C(R)|}{|S(R)|}$$

Cette fois, on se demande plutôt quelle est, parmi les documents choisis, la proportion de « bons » documents. Le taux précision ***Pr(R)*** a pour dual le taux de *bruit Br(R)* : par définition, *Br(R) = 1 - **Pr(R)***.

En pratique, il y a pratiquement toujours une tension entre le rappel et la précision. L'amélioration sur un plan tend à dégrader la performance sur l'autre plan, et on recherche généralement un compromis acceptable.

Le programme TIPSTER de DARPA (une agence de la défense américaine) comprend un volet recherche documentaire qui donne lieu à une concours amical appelé *Text Retrieval Conferences* (TREC) qui permet à tous ceux qui le désirent de mesurer devant diverses tâches de recherche documentaire. Durant les premières années, les tâches en question ne couvraient que la langue anglaise, mais la tendance est maintenant à couvrir d'autres langues, soit dans des tâches séparées, soit dans des tâches multilingues. Bien que le TREC ne comprenne pas de concours à grande échelle portant sur la langue française, plusieurs équipes de la francophonie y ont participé sur les tâches anglaises ou multilingues.

De plus, dans le cadre de ses actions de recherche concertées, l'AUPELF/UREF a récemment mis en place le projet Amaryllis qui permet aux groupes de la francophonie de se donner leur propre concours amical, sur un modèle similaire à celui des TREC. À noter toutefois que contrairement aux TREC, la participation du secteur privé à Amaryllis est demeurée jusqu'à maintenant plutôt erratique, peut-être en raison du caractère plutôt universitaire des structures mises en place.

2.4. Traitement linguistique : extraction automatique d'information, résumeur automatique, traduction automatique

L'information circule généralement sous forme de documents dont le contenu est la plupart du temps du texte en langue naturelle. De plus en plus, ces documents sont

produits, distribués et stockés en format électronique. À chacune de ces étapes, on peut faire intervenir des outils informatiques pour faciliter l'une ou l'autre des tâches qui sont en cause.

L'étape de production comprend au premier chef les tâches de rédaction et de traduction, cette dernière pouvant aussi se retrouver à l'étape de consommation.

1.1.4. Génération automatique et aides à la rédaction

Dans certains cas, il est possible de générer automatiquement des textes qui répondent à certains besoins précis. Au nombre de ces cas, on compte notamment ceux où les textes à produire constituent des descriptions de données objectives.

Génération automatique

Par exemple, Kukich 4 a montré qu'il était possible de construire un système capable d'analyser les fluctuations des cotes boursières pour en extraire l'information jugée importante, puis de produire automatiquement un texte qui verbalise cette information. Et Kittredge 4 a montré qu'il était possible de générer automatiquement, à partir des données objectives normalement fournies à un rédacteur humain, des bulletins météorologiques dans deux langues différentes. Ainsi, la génération automatique peut donc à l'occasion permettre de court-circuiter aussi bien l'étape de la traduction (humaine ou automatique) que l'étape de la rédaction humaine.

Il n'en reste pas moins que, dans l'écrasante majorité des cas, les textes doivent être produits par des humains, qu'ils soient ou non des rédacteurs professionnels. La machine peut alors être exploitée pour fournir différentes formes d'aide à la tâche de rédaction.

Dictionnaires et grammaires en ligne

La plupart des rédacteurs utilisent déjà un système informatisé de traitement de texte ou d'éditique pour effectuer la saisie de leur texte. Beaucoup de rédacteurs font également appel à des ouvrages de référence en ligne: dictionnaires, grammaires, conjugueurs, etc. Il ne s'agit pas ici d'un traitement informatisé de la langue, mais plus simplement de l'informatisation de ressources linguistiques traditionnelles. Bien que l'effort technologique requis pour une telle informatisation paraisse plutôt modeste, il n'en reste pas moins que le catalogue des ouvrages de référence linguistiques informatisés du français est encore bien incomplet. Et lorsque les outils sont disponibles, leur niveau d'intégration avec les systèmes de traitement de texte et d'éditique populaires laisse souvent à désirer.

Correcteurs

Pour ce qui est des outils fondés sur le traitement automatique de la langue, on constate qu'un nombre croissant de rédacteurs ont recours à des outils d'aide à la relecture, comme les vérificateurs d'orthographe et de grammaire. La langue française n'est pas en

reste sur ce plan puisqu'il existe déjà sur le marché plusieurs correcteurs du français dont la performance se compare très bien à ceux qui existent pour l'anglais[1].

On doit cependant garder à l'esprit que les technologies existantes sont loin d'être parfaites. Non seulement leur arrive-t-il de laisser passer des erreurs (leur « rappel » n'est pas parfait), mais ils ont aussi une nette tendance à détecter des « erreurs » qui n'en sont pas (leur « précision » n'est pas parfaite). Leur utilisateur doit alors être en mesure de faire la part des choses. Par conséquent, les systèmes existants sont plus utiles pour aider un rédacteur compétent à repérer ses « erreurs d'inattention » que pour combler des lacunes sérieuses dans la compétence linguistique d'un rédacteur.

Certaines recherches sont en cours pour mettre au point des systèmes d'aide à la frappe 4. Compte tenu du contexte créé par une portion de texte déjà frappée, un modèle de langue peut souvent permettre d'anticiper un fragment de la suite désirée. Typiquement, la prédiction concernera le reste du mot en train d'être frappé et/ou un petit nombre de mots à sa suite. Par exemple, si je tape « au fur... », le système pourra normalement anticiper « ...et à mesure ». Dans le cas d'un très bon dactylo, il n'est pas certain que cette approche soit viable, car il pourrait s'avérer plus long pour lui de valider les prédictions du système que de taper tout le texte lui-même. Par contre, pour les dactylo moins rapides, et surtout pour ceux qui ont des handicaps qui entravent le processus de frappe 4, cette approche peut s'avérer très profitable.

Réaccentuation automatique

Ce n'est que tout récemment qu'ont commencé à apparaître sur l'Internet des standards qui permettent d'échanger des textes encodés avec des jeux de caractères plus complets que le fameux ASCII (voir section 2.2) et bon nombre de francophones sont encore aux prises avec des situations où il est difficile ou risqué d'utiliser les diacritiques normaux de leur langue, par exemple dans leur courriel. Un texte français parvient donc souvent à son destinataire dans une forme privée d'accents. Il existe aussi de masses de textes qui ont été préparées à une époque où il paraissait inévitable, voire normal, que des textes stockés sur ordinateur (par exemple une base de données de résumés de textes) soient dépourvus d'accents.

La perte des caractères accentués entraîne des inconvénients sérieux. D'abord, pour un texte un moindrement long, le lecteur ne peut compenser cette perte qu'au prix d'un effort cognitif supplémentaire qui devient vite très désagréable. Deuxièmement, cette perte compromet sérieusement la possibilité d'effectuer avec succès des traitements linguistiques comme la traduction automatique ou la synthèse de parole[2]. Et troisièmement, l'utilisation d'un texte sans accents dans un contexte plus soigné, par exemple une publication, est généralement proscrite.

La réinsertion manuelle des accents dans un texte français qui en est dépourvu est une tâche particulièrement fastidieuse. Fort heureusement, il est maintenant démontré que les

[1] Et ce, en dépit du fait que la langue française comporte souvent des difficultés plus grandes à ce niveau, comme c'est le cas pour les règles d'accord.

[2] On notera la conséquence de ce dernier fait pour les malvoyants. S'ils peuvent généralement utiliser une machine qui leur lit un texte à voix haute, ceci devient pratiquement impossible dans le cas d'un courriel sans accents.

systèmes de réaccentuation automatique peuvent atteindre une très bonne performance en français. La plupart du temps, le lecteur pourra facilement tolérer les quelques erreurs effectuées par le système, et dans le cas où on a besoin d'un texte parfait la révision de ces erreurs constituera une tâche relativement légère.

1.1.5. Traduction automatique et outils d'aide au traducteur

La traduction automatique

Historiquement, la traduction automatique (TA) est l'application qui a suscité les premières recherches en traitement automatique des langues naturelles, dès le début des années cinquante. Si certains chercheurs ont pu à l'époque croire que l'émergence d'une machine à traduire universelle était imminente, ce ne fut pas le cas de Yehoshua Bar-Hillel. Dans un article intitulé *The State of Machine Translation in 1951* 4, celui-ci observait qu'en l'absence de méthodes générales pour résoudre les « ambiguïtés sémantiques », les systèmes de traduction automatique (TA) ne pouvaient pas produire des traductions de bonne qualité. Quarante-six ans plus tard, nous sommes bien forcés d'admettre que son pessimisme était pleinement justifié.

L'objectif ultime de construire une machine capable de rivaliser avec le traducteur humain n'a pas cessé de fuir par devant les lentes avancées de la recherche. Les approches traditionnelles *à base de règles* ont conduit à des systèmes qui tendent à s'effondrer sous leur propre poids bien avant de s'élever au-dessus des nuages de l'ambiguïté sémantique. Les approches récentes *à base de corpus* (qu'elles soient fondées sur les méthodes statistiques ou les méthodes analogiques) promettent bien de réduire la quantité de travail manuel requise pour construire un système de TA, mais il est moins sûr qu'elles promettent des améliorations substantielles de la qualité des traductions machine.

Le pessimisme de Bar-Hillel face à l'objectif ultime de la TA était contrebalancé par un certain optimisme quant à la possibilité mettre la technologie à contribution en dépit de toute son imperfection. À cette fin, il proposait déjà trois stratégies d'application qui demeurent encore aujourd'hui le lot de ceux qui se tournent vers la TA.

La première stratégie mise de l'avant consistait à utiliser la TA comme une aide à ceux qui ont besoin de « balayer l'immense production écrite [d'ennemis réels ou potentiels] », en leur fournissant des traductions approximatives. Une telle traduction peut en effet aider un analyste à déterminer si un certain document en langue étrangère est potentiellement intéressant pour ses fins. Au besoin, les documents jugés intéressants pourront ensuite être envoyés à un traducteur humain.

Historiquement, c'est cette utilisation comme un *outil d'assimilation d'information* qui a constitué le débouché commercial le plus important de la TA. Et si le client traditionnel se trouvait surtout au sein des « agences de renseignements », l'Internet semble en voie de créer une demande populaire pour des outils de repérage d'information multilingue. C'est ce nouveau marché que vise par exemple le projet ATS de la société Alis Technologies.

La deuxième stratégie que Bar-Hillel proposait consistait à cibler des situations de « langue restreinte ». En 1977, l'implantation réussie du système MÉTÉO 4, qui traduit

les prévisions météorologiques du ministère de l'Environnement canadien, confirma que la TA peut fonctionner très bien lorsqu'elle est appliquée à des *sous-langues naturelles* simples. Malheureusement, il semble y avoir tellement peu de situations de ce genre sur le marché de la traduction que cette victoire de la TA est demeurée un fait plus ou moins isolé. La stratégie apparentée qui consiste à imposer des contraintes externes sur le processus de rédaction en vue de faciliter la TA fait présentement l'objet de certains efforts (par exemple le système KANT développé par l'Université Carnegie-Mellon pour la société Caterpillar), mais sa viabilité n'a pas encore été démontrée de façon convaincante.

Troisièmement, dans les cas où on a besoin de traductions de haute qualité de textes complexes, Bar-Hillel estimait qu'on pouvait coupler l'humain et la machine d'une manière efficace. L'humain aurait pour rôle soit de prévenir les erreurs de la machine en annotant le texte avant de le passer à la machine (« préédition ») soit de les corriger après coup (« postédition »). L'avènement d'ordinateurs plus puissants eut pour effet d'ajouter une autre possibilité : celle d'une interaction homme/machine pendant le processus de TA.

Malheureusement, il est maintenant évident que Bar-Hillel avait sous-estimé le problème d'une division du travail acceptable et efficace entre l'humain et le système de TA. On a certes régulièrement entendu des affirmations voulant qu'il en coûte beaucoup moins de postéditer les sorties de systèmes comme SYSTRAN, WEIDNER, LOGOS et METAL que de traduire par les méthodes classiques. Toutefois, des expériences contrôlées de manière rigoureuse ont montré que dans certains cas le contraire était vrai (pour un bon exemple, voir 4). Ces conclusions négatives s'accordent mieux avec le fait indiscutable qu'en dépit d'efforts très importants, la pénétration de la TA au sein des services de traduction professionnels demeure au mieux marginale.

Les trois stratégies proposées par Bar-Hillel épuisent l'ensemble des compromis logiquement possibles par rapport aux attributs du système de TA idéal (qualité, généralité, automatisation complète). Par conséquent, d'ici à ce qu'une percée imprévue ne survienne, l'utilisation de la TA demeurera plus ou moins limitée aux tâches d'assimilation d'information (repérage) ou aux tâches de dissémination de textes relevant de sous-langues restreintes.

Les outils d'aide au traducteur

Dans la plupart des situations où l'on a besoin de traductions de haute qualité, la machine doit savoir garder sa place, et cette place demeure comme la décrivait Martin Kay 4 il y a plus de quinze ans[1] :

> *« Je veux préconiser une approche selon laquelle on permettrait à la machine de prendre en charge graduellement, presque imperceptiblement, certaines fonctions du processus global de traduction. La machine assumerait d'abord des tâches périphériques. Puis, peu à peu elle s'attaquerait au cœur du processus. La démarche serait toujours empreinte de modestie. Jamais n'essaierait-on de faire plus que ce que l'on sait bien faire.»*

[1] L'article que nous citons vient tout juste d'être publié, mais il circule sous forme miméographiée depuis 1980.

Cette démarche s'oppose à la TA parce que son caractère graduel implique une technologie qui soit capable de se plier à la démarche naturelle du traducteur humain. On ne construit pas un traducteur robot, mais plutôt un poste de travail pour le traducteur humain.

Si cette idée a tout de suite suscité la sympathie dans les milieux de recherche, elle est longtemps restée lettre morte, sans doute parce que le paradigme scientifique qui était alors en vogue (les systèmes à base de règles) se prêtait plutôt mal à sa réalisation. L'émergence récente des approches à base de corpus vient heureusement de donner un élan important à l'approche poste de travail.

Les recherches récentes sur les méthodes probabilistes ont en effet permis de démontrer qu'il était possible de modéliser d'une manière extrêmement efficace certains aspects simples du rapport traductionnel entre deux textes. Par exemple, on a mis au point des méthodes qui permettent de calculer le bon « alignement » entre les phrases d'un texte et de sa traduction, c'est-à-dire d'identifier à quelle(s) phrase(s) du texte d'origine correspond chaque phrase de la traduction. Appliquées à grande échelle, ces techniques permettent de constituer, à partir des archives d'un service de traduction, une *mémoire de traduction* qui permettra souvent de recycler des fragments de traductions antérieures. Des systèmes de ce genre ont déjà commencé à apparaître sur le marché (Translation Manager II de IBM, Translator's Workbench de Trados, TransSearch, etc.).

Les recherches les plus récentes se concentrent sur des modèles capables d'établir automatiquement les correspondances à un niveau plus fin que celui de la phrase : syntagmes et mots. Les résultats obtenus laissent entrevoir toute une famille de nouveaux outils pour le traducteur humain, dont les aides au dépouillement terminologique, les aides à la dictée et à la frappe des traductions ainsi que les détecteurs de fautes de traduction.

1.1.6. L'aiguillage de documents

Le système de distribution traditionnel se bornait à acheminer au consommateur des ensembles de documents (par exemple un journal) assemblés par le producteur lui-même. La pléthore d'information qui affecte un nombre croissant de personnes est en train de faire éclore une couche de « courtiers » entre les fournisseurs et les consommateurs d'information. Les services de « veille » et les services Internet de type « poussée » proposent à leurs clients une information déjà filtrée en fonction de leurs intérêts particuliers.

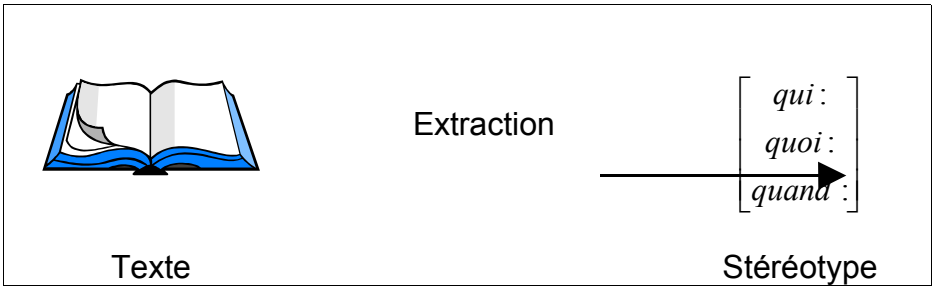
Le courtier en documents doit effectuer leur aiguillage en fonction du profil d'intérêts particulier à chaque client. Chaque client du courtier lui délègue en quelque sorte une tâche particulière de *filtrage* de documents et les techniques de base seront donc celles du filtrage (voir section 2.3).

1.1.7. Extraction automatique d'information

Dans certains cas, une technologie capable d'identifier quels sont les documents pertinents ne suffit pas, car le nombre de ces documents est tellement élevé qu'il est impossible de les lire tous.

Si la structure de l'information recherchée est suffisamment simple, on peut alors se tourner vers des technologies de « lecture automatique », plus modestement appelées technologies d'*extraction automatique d'information*.

Typiquement, on s'intéresse à des situations où l'information à extraire des textes prendra la forme d'enregistrements spécifiques dans une base de données. Le contenu des textes visés n'est pertinent que dans la mesure où il permet de remplir les cases d'un stéréotype :



Par exemple, on pourrait s'intéresser à construire un système capable de traiter les pages financières d'un ou plusieurs journaux de manière à en extraire une description précise de chaque cas de fusion ou d'acquisition de sociétés commerciales mentionné dans ces pages. On construirait ainsi une base de donnée contenant des enregistrements du genre :

$$\begin{bmatrix} \text{Opération} = \textit{Achat} \\ \text{Par} = \textit{ABC inc.} \\ \text{De} = \textit{XYZ inc.} \\ \text{Montant} = 2\,000\,000\,\$ \\ \text{Date} = 07/09/97 \\ \text{etc.} = \textit{etc.} \end{bmatrix}$$

Il est très facile d'imaginer des applications pour ce genre de technologie, et il est évident que la défense américaine s'y intéresse beaucoup depuis quelques années. En fait, le programme TIPSTER de l'agence DARPA mentionné plus haut comprend aussi un volet extraction d'information qui a lui aussi donné lieu à une série de concours amicaux, sous le nom de *Message Understanding Conferences* (MUC).

Contrairement à la situation décrite plus haut en recherche documentaire, il semble que: 1) aucun laboratoire de la Francophonie n'a participé aux évaluations MUC passées; et 2) il n'existe aucun concours amical dans ce domaine au sein de la Francophonie. Il ne paraît donc pas exagéré de dire que la Francophonie est en train de prendre du retard dans les technologies d'extraction d'information.

1.1.8. Résumeur automatique

Malheureusement, dans beaucoup de cas où la masse des documents jugés pertinents est trop grande, l'extraction automatique d'information ne s'avère pas réalisable. Dans ces cas, la machine pourra souvent faciliter quand même la tâche de l'humain en lui présentant une version résumée ou condensée de chaque texte – voire une version résumée ou condensée de la masse de textes dans son ensemble.

L'idéal serait bien sûr de disposer d'une technologie capable de résumer l'information de la même façon qu'un rédacteur ou un documentaliste humain pourraient le faire. Dans l'état actuel de la technologie linguistique, ceci n'est malheureusement pas possible. La production d'un bon résumé constitue en effet l'un des défis les plus difficiles que l'on puisse imaginer pour la technologie de TALN. Cette tâche met systématiquement et simultanément à l'épreuve : a) une compétence à comprendre des textes en profondeur ; b) une compétence à juger de l'importance relative des éléments de leur contenu (ce qui implique une compréhension approfondie du référentiel extralinguistique de ces textes) ; et c) une compétence à rédiger un nouveau texte, à savoir le résumé.

À défaut de vrais résumés, il paraît cependant possible d'extraire d'un texte complet les portions qui paraissent les plus importantes, de manière à en obtenir une version « condensée ». Les portions en question seront généralement des paragraphes ou des phrases. À cette fin, on pourra utiliser diverses techniques. L'approche privilégiée par plusieurs chercheurs consiste à transposer à ce problème les méthodes de la recherche et du filtrage d'information, comme évoqué ci-dessus. Soit un document représenté comme un vecteur de mots pondérés :

$$D : \{ (mot_1, poids_{1,1}), (mot_2, poids_{2,1}), \ldots (mot_n, poids_{n,1})\}$$

On peut représenter chacune des portions de *D*, disons les phrases de *D*, de manière analogue :

$$P_1 : \{ (mot_1, poids_{1,1}), (mot_2, poids_{2,1}), \ldots (mot_n, poids_{n,1}) \}$$
$$P_2 : \{ (mot_1, poids_{1,1}), (mot_2, poids_{2,1}), \ldots (mot_n, poids_{n,1}) \}$$
$$\ldots$$
$$P_n : \{ (mot_1, poids_{1,1}), (mot_2, poids_{2,1}), \ldots (mot_n, poids_{n,1}) \}$$

Ensuite, on mesure la similarité entre *D* et chaque P_i, et on retient les phrases qui ressemblent le plus à *D*.

2.5. Reconnaissance et synthèse vocales

La parole constitue pour l'être humain le moyen de communication le plus naturel, le plus efficace et le plus aisé. C'est pour cette raison que beaucoup de fabricants d'appareils électroniques (y compris les ordinateurs) s'y intéressent comme interface entre leurs produits et les utilisateurs. On a vu récemment de nombreux exemples d'applications des technologies de la parole : messagerie unifiée (messagerie vocale, courriel et télécopie dans une seule boîte à lettre, accessible par téléphone), système de réponse vocale

interactif, commande vocale d'ordinateurs ou autres appareils, formation linguistique, ordinateurs parlants, dictionnaires et traducteurs électroniques portatifs, composition mains libres de téléphones cellulaires, navigation par la voix.

Les technologies de la parole (reconnaissance vocale, synthèse vocale et compression de son numérique) ne sont pas nouvelles ; le nombre d'applications et de produits augmente, mais jusqu'à récemment leur diffusion a été limitée par des performances insuffisantes et des prix trop élevés. La reconnaissance vocale exigeait traditionnellement un entraînement de l'appareil pour lui faire reconnaître un nombre limité de mots énoncés par un seul locuteur ; de surcroît, la sensibilité au bruit ambiant était excessive, de même que la sensibilité aux variations d'élocution dues à la fatigue, à la tension ou à la maladie du locuteur. Ces systèmes étaient basés sur les mots et exigeaient une élocution discontinue, mot par mot, au lieu de la parole continue naturelle.

La synthèse vocale souffrait de même de défauts liés au manque de raffinement des algorithmes de synthèse et au manque de puissance de calcul et de mémoire à prix abordable. Le résultat était des systèmes chers, au son monotone, robotique et avec de nombreuses erreurs de prononciation (prosodie et liaisons notamment).

En plus des problèmes de prix et de performance, le développement de la reconnaissance et de la synthèse vocale en de nouvelles versions linguistiques s'est avéré long et coûteux, avec pour résultat que la plupart des développement se sont concentrés sur une seule langue (généralement l'anglais) ou au mieux sur quelques-unes. Il est clair qu'une telle situation ne favorise ni le français ni ses langues partenaires.

3. Stratégie

1.2. Évolution

Il est à prévoir que le traitement informatique du langage évoluera rapidement, en phase avec le développement général de l'informatique et stimulé par la découverte de nouvelles applications utiles et largement commercialisables. Une étude du groupe Gartner prétend que « *En 1999, la compréhension de la langue naturelle ne sera plus une niche* » 4, ce qui signifie tout simplement que de nombreuses applications de grande diffusion utiliseront cette technologie, qui relève du domaine plus large du traitement informatique du langage.

Il faut maintenant distinguer d'une part l'évolution du domaine en général de celle prévisible pour le français et ses langues partenaires, et d'autre part l'évolution des différents sous-domaines du traitement informatique. Nul ne s'étonnera qu'une grosse fraction du travail fait dans ce domaine concerne la langue anglaise, pour des raisons évidentes : domination américaine en informatique en général, marché gigantesque des pays anglo-saxons, et effet d'entraînement dû à la présence d'une masse critique (on dit que 80 % des données informatiques dans le monde sont en anglais 4, et Internet est pareillement dominé 4). Il suffit par exemple de recenser les paires de langues disponibles en traduction automatique (ou assistée par ordinateur) pour constater que l'anglais y figure presque toujours, même quand les logiciels sont développés hors du monde anglo-saxon.

Malgré cette domination de l'anglais, une évolution prévisible qui favorise en fait le français et toutes ses langues partenaires est le passage graduel au JUC comme fondement du codage du texte dans toutes les applications. Étant donné les problèmes de conversion qu'une telle migration impose et l'inertie des marchés (parc installé, etc.), l'avènement du JUC sera certainement assez lent et peut-être aussi semé d'embûches. La Francophonie a intérêt à favoriser ce changement, mais doit aussi se prémunir contre les problèmes qu'il causera : compatibilité de données et de logiciels, lenteur de migration, logiciels non adaptés.

1.2.1. Traduction automatique

Les systèmes de traduction automatique vont devenir de plus en plus accessibles à tous, couvrant de plus en plus de langues, bien que cette couverture se dessinera le plus souvent en étoile autour de la langue anglaise. La qualité des traductions machines ne s'améliorera que très lentement. Par conséquent, la traduction automatique demeurera plutôt un outil de veille et d'assimilation d'information qu'un outil de diffusion. Dans cette fonction d'assimilation, on verra aussi se répandre des outils d'aide à la lecture en langue seconde (par exemple, aide à la lecture en anglais à l'intention des francophones). Les pouvoirs publics et les entreprises continueront à recourir à des traducteurs humains, mais ceux-ci bénéficieront de l'émergence de nouvelles aides comme les mémoires de traduction ou les vérificateurs de traduction.

1.2.2. Recherche et indexage

Il est évident que les interfaces en langue naturelle sont appelées à jouer un rôle de plus en plus important, à mesure que de plus en plus d'utilisateurs, de moins en moins spécialistes du domaine documentaire, voudront trouver de l'information sur l'Internet et les futures inforoutes ou dans leurs intranets. Il s'agit donc d'un secteur d'avenir auquel la Francophonie devrait s'intéresser.

De telles interfaces offrent plusieurs avantages : le premier, évident, est qu'elles permettent une recherche efficace par des personnes non spécialistes du domaine de la recherche documentaire. En seconde analyse, il apparaît que des logiciels capables de « comprendre » une question en langue naturelle seront aussi à même d'interpréter et de « comprendre » les documents du fond auxquels ils accèdent, ouvrant ainsi la porte à des recherches beaucoup plus efficaces et fructueuses. Ainsi, lors d'une recherche sur « le manque d'éthique dans le milieu politique », un logiciel intelligent pourra trouver pertinent un document mentionnant que « le maire de X a touché un pot-de-vin », ce qui est tout à fait hors de portée d'une recherche par mots-clés ou par signature[1]. Il s'agit ici d'une amélioration du rappel (trouver ce que l'on cherche) ; la précision de la recherche (ne trouver **que** ce qu'on cherche) fera sans aucun doute l'objet d'une évolution favorable sous l'impulsion des technologies d'ingénierie langagière.

Dès lors, une extension évidente est le multilinguisme. Le même logiciel intelligent ne pourrait-il pas trouver un document contenant « the mayor of X pocketed a backhander » ? La présence sur l'Internet de documents en multiples langues, et même parfois multilingues, rend cette possibilité extrêmement attrayante.

Le projet SPIRIT de la société française T.Gid[2] s'attaque à ce problème, allant même jusqu'au multilinguisme. Nous ne pouvons qu'approuver cette démarche, mais devons toutefois mentionner deux aspects qui nous semblent limitatifs : le serveur SPIRIT-W3 est accessible de l'Internet, permettant au monde d'interroger sa base de données, mais il ne semble pas que ses capacités d'analyse linguistique puissent être utilisées pour fouiller l'Internet. C'est donc une voie à sens unique. D'autre part, le multilinguisme est fort limité au départ, ce qui se comprend aisément, mais il appert que le manque d'attention dès l'origine aux problèmes de jeux de caractères pourrait nuire considérablement à l'extension future à d'autres langues ; il y aura problème dès que l'ISO Latin-1 ne suffira plus. L'Internet est farci d'exemples de logiciels et de protocoles, par ailleurs très bons, qui se butent à ce problème en voulant s'étendre dans le monde. Alis a d'ailleurs dans ses cartons un projet intéressant plusieurs partenaires majeurs, projet reprenant d'une part une technologie linguistique plus avancée, et d'autre part une base multilingue fondée sur Unicode permettant d'attaquer toutes les langues dans le domaine de la recherche documentaire.

1.2.3. Résumeurs

La simple intuition nous suggère que la technologie de génération automatique de résumés, si elle venait à être perfectionnée à un niveau acceptable, pourrait connaître un

[1] Voir la section 2.3, Recherche et filtrage d'information.

[2] Page d'accueil à http://www.refer.fr:8001/

essor considérable. Il est parfaitement clair qu'une telle technologie répond à un besoin bien précis et très répandu, à savoir celui de diminuer l'« infobésité » que nous impose la société de l'information : il est beaucoup plus facile et efficace de lire et de trier des résumés brefs et précis que de fouiller la masse entière de documentation. Les résumés suffisent dans bien des cas, et permettent presque toujours de séparer l'ivraie du bon grain. D'autre part, on peut espérer que les textes issus des résumeurs, ayant été créés par des générateurs automatiques, soient dans une langue plus contrôlée et ainsi plus apte à la traduction automatique[1]. Étant donné l'état de l'art en traduction automatique, on peut donc envisager qu'un document traduit soit peu ou pas lisible, alors que son résumé, traduit lui aussi, le soit ! Le domaine des résumeurs est donc porteur de par ses applications possibles, et digne d'intérêt.

3.1.1. Veille documentaire

Les activités de veille technologique – en réalité de veille documentaire en général – profiteront au premier chef des améliorations prévisibles des résumeurs et des moteurs de recherche, permettant d'avoir une prise sur la masse grandissante de documentation en ligne sur les inforoutes et dans les intranets. La traduction automatique aura aussi un impact positif ; en fait, un certain type bien particulier de veille, l'espionnage, est à l'origine de beaucoup des développements originaux dans ce domaine.

Le filtrage d'information aura aussi un impact, permettant un type de veille passive (du point de vue du veilleur) par le biais de « journaux sur mesure », personnalisés et utilisant les technologies de « poussée ». L'utilisateur, qui n'est plus nécessairement un professionnel de la veille, s'abonne au service, règle le filtrage en fonction de ses champs d'intérêt et de l'information qu'il veut surveiller, et reçoit par la suite (via la « poussée ») toute information, toute nouvelle qui a franchi son filtre personnalisé.

1.3. Risques

On peut identifier un certain nombre de facteurs de risque pouvant affecter le traitement informatique des langues française ou partenaires. D'une part, il importe d'être bien conscient de l'importance grandissante et déjà majeure des technologies de l'information, comme en témoignent les paroles du vice-président des États-Unis :

> *«Aujourd'hui plus que jamais, nos entreprises vivent d'informations. Un réseau de l'information rapide et souple est tout aussi vital à nos industries que le plastique et l'acier.*
>
> *Si nous n'agissons pas de manière décisive pour nous assurer que les États-Unis auront l'infrastructure de l'information dont ils ont besoin, chaque entreprise, chaque consommateur américain en pâtira...»*

- Al Gore, v.-p. des É.U., devant le National Press Club le 21 XII 1993

[1] On parle ici des vrais résumeurs, encore à l'étape théorique, et non pas des simples condenseurs qui extraient des phrases pertinentes du texte.

On voit ici poindre le premier risque, et sans doute le principal, qui menace la Francophonie : la marginalisation. Les Américains – et bien d'autres – sont tout à fait conscients de l'importance de ces technologies, et entendent les dominer. Or ils en ont les moyens, bénéficiant du premier rang et du pouvoir économique et politique que l'on sait, et il est à prévoir qu'ils chercheront à établir et maintenir leur domination sans égard aux langues de la Francophonie.

On peut distinguer deux aspects à ce risque : d'une part, on peut craindre que les produits disponibles sur le marché soient moins bien adaptés au français et à ses langues partenaires, c'est-à-dire qu'ils fonctionnent moins bien, souffrent de plus de limitations, soient disponibles plus tard ou à coût plus élevé que leurs équivalents en anglais. Dans les domaines en émergence, on peut par exemple penser à la pauvreté relative de la qualité des logiciels de traduction automatique du ou vers le français (sauf quand l'autre langue est l'anglais) ou même à leur absence : l'homme d'affaire francophone qui reçoit un message ou consulte une page Web dans une langue X ne pourra pas traduire dans sa langue rapidement et à peu de frais, alors que son concurrent anglophone le pourra. De même, les logiciels de reconnaissance vocale développés d'abord pour l'anglais risquent d'être moins bien adaptés au français, plus sensibles à l'accent du locuteur, et donc moins utiles. Et qui pourra profiter des résumeurs automatiques ? Étant donné l'omniprésence des technologies de l'information, une telle situation ne peut manquer de nuire très sérieusement à un grand nombre de secteurs de l'économie, menant à une position concurrentielle désavantageuse des acteurs de la Francophonie.

À défaut de relever avec succès le défi de la création et de la commercialisation d'outils de repérages multilingues, il y a fort à parier que le peu de contenu non anglophones sur les inforoutes perdra à l'échelle internationale le peu de visibilité qu'il lui reste au profit de textes, de bandes sonores ou de vidéos repérables à l'aide des outils américains disponibles ou en voie d'élaboration. Les mêmes effets néfastes de cette carence d'outils efficaces se feraient sentir dans les entreprises où des outils de repérage surannés au sein de l'intranet se révéleraient moins efficaces que ceux disponibles aux concurrents anglophones, avec comme conséquences ou bien une perte de compétitivité, ou bien une pression vers l'utilisation de l'anglais comme langue de travail unique.

L'autre aspect concerne directement les secteurs des technologies de l'information. Ici, la concurrence est rendue plus ardue par un phénomène propre aux logiciels, à savoir le besoin de compatibilité. Le fabricant d'un grille-pain ne doit s'inquiéter que de la compatibilité de son appareil avec la prise de courant, mais le concepteur d'un logiciel doit s'assurer qu'il fonctionne bien avec tous les autres logiciels avec lesquels il devra coopérer, à commencer par le système d'exploitation et sans oublier les jeux de caractères, formats de fichier, protocoles de réseau, et ainsi de suite. Or la compétition est nettement plus rude lorsque les concurrents contrôlent les normes, que ce soit *de jure* ou *de facto* ; en d'autres termes, la domination des marchés s'autoentretient, le dominateur ayant l'occasion de façonner le marché de manière à assurer la pérennité de sa domination. Il est donc vital de ne pas laisser ce cercle vicieux s'installer au détriment de la Francophonie.

3.2. Actions à prendre, priorités pour la Francophonie

Nous avons vu dans la section précédente qu'il y avait à la fois possibilité et désir – bien naturel dans un marché concurrentiel – d'établissement d'une hégémonie, principalement américaine, mais en fait bénéficiant à tous les pays anglophones, dans le domaine des technologies de l'information. Cette hégémonie peut s'établir – et en fait est déjà assez bien établie – par le simple jeu de la concurrence aidée par la nature des besoins de compatibilité des logiciels ; nul besoin d'intentions machiavéliques ou de manœuvres malhonnêtes.

La première priorité pour la Francophonie devrait donc être de tenter de contrer cette domination qui lui est désavantageuse, de façon à éviter sa marginalisation et les conséquences fâcheuses qui en découlent. La nature du problème impose la nécessité d'actions concertées ; sa gravité ne laisse pas place aux états d'âme ou aux arrière-pensées qui diminueraient l'efficacité des actions impérieuses à prendre dans un contexte de globalisation et de concurrence féroce.

Dans ce combat, car s'en est un, à peu près toutes les langues peuvent être considérées comme partenaires ; la même menace pèse en effet sur toutes et bien peu peuvent prétendre avoir la masse critique les mettant à l'abri de la marginalisation. La prise de conscience du problème est toutefois fort variable, les locuteurs de certaines langues s'objectant peu à l'établissement de l'anglais comme *lingua franca* et ne percevant pas les avantages que ce statut confère à certains de leurs compétiteurs, et ce bien sûr à leur détriment.

Deux sous-domaines des technologies de l'information viennent à l'esprit, où des actions concertées de la Francophonie seraient importantes et pourraient être efficaces. Premièrement, le traitement automatique du langage est un domaine en émergence qui, d'une part, concerne directement la langue, et qui, d'autre part, est prometteur d'applications nombreuses et économiquement importantes. Deuxièmement, nul ne peut plus nier l'importance des autoroutes de l'information, en particulier de l'incontournable Internet et de l'application de ses techniques à l'intérieur des entreprises, à savoir les intranets. Voici donc quelques suggestions d'actions touchant principalement ces deux domaines :

3.2.1. Favoriser la normalisation en français, pour le français

La normalisation joue un grand rôle dans les technologies de l'information, pour les raisons données à la section 3.2, Risques. Il est désolant de constater que les activités de normalisation dans ce domaine tiennent fort peu compte des besoins du français (et encore moins des langues partenaires à écriture autre que latine), et qu'elles ne se déroulent que très rarement en français. L'ISO[1] et la CÉI[2] ont toutes deux le français, l'anglais et le russe comme langues officielles ; toutefois, la normalisation en technologies

[1] **Organisation internationale de normalisation** en français, **International Organization for Standardization** en anglais, **Международная Организация по Стандартизации** en russe. L'abréviation « ISO » est dérivée du grec *isos*, signifiant « égal ».

de l'information de ces deux organisations est confiée au Comité conjoint 1 (JTC1) qui ne fonctionne, en pratique, qu'en anglais. Il conviendrait d'encourager et de favoriser la publication *simultanée* des versions française et anglaise des normes ISO et CÉI. Les longs retards observés, dans le meilleur des cas, obligent les acteurs à utiliser les versions anglaises pour rester à jour, au détriment de la dissémination de la terminologie française dans ces domaines.

Nous pensons qu'il serait souhaitable de renforcer la présence francophone dans les forums de normalisation de terminologie suivants (entre autres) : ISO/CÉI JTC1/SC1 (aménagement terminologique des technologies de l'information) et ISO/CÉI TC37 (outils terminologiques). La terminologie est importante en ce qu'elle sert de base pour les lexiques (eux-mêmes utilisés pour des applications comme le glosage[1]), pour la traduction et les aides à la traduction, etc.

Les normes ISO coûtent très cher ; d'autre part, certains standards, notamment les standards Internet, continueront à n'être publiés qu'en anglais dans un avenir prévisible. En conséquence, nous recommandons que les normes et standards d'importance soient publiés sous forme de commentaire sur la norme (traduite, en préparation de traduction ou encore non traduite) et de rendre ce commentaire disponible à très bas prix ou gratuitement, sous la forme de cédéroms ou sur l'Internet. Certaines normes comme l'ISO 10646 représentent en effet un fondement de la nouvelle informatique multilingue, elles devraient être disponibles depuis longtemps au plus grand nombre en français. En matière terminologique également, l'ISO 10646 représente une mine de termes extrêmement importants : elle reprend les noms de tous les caractères (lettres, signes et symboles) du Jeu universel de caractères. Il faut exploiter ce gisement et faire en sorte que les noms français de ces caractères soient disponibles de par la Francophonie.

Il faut de plus insister sur une plus grande participation de francophones aux différentes instances de normalisation (autres que nationales), pour favoriser la normalisation **pour** le français, ou du moins ne le négligeant pas. Il suffit de se souvenir de l'incident du codage Latin-1 sans **œ**, **Œ** et **Ÿ**, ou celui plus récent au cours duquel le Latin-0 – qui cherche rappelons-le à corriger les carences du Latin-1 pour le français – fut battu en brèche lors d'une réunion de normalisation européenne où les francophones brillaient par leur absence, alors que les Américains étaient présents en force.

Il faut donc une plus grande présence dans ces instances de normalisation, sans négliger celle qui n'ont pas caractère officiel (ISO, CÉI, etc.) mais qui n'en ont pas moins une influence cruciale sur la pratique. Par exemple, il ne serait pas mauvais de s'assurer que les modèles d'objets répartis (Corba, du Object Management Group) et les nouvelles spécifications du W3C[2] respectent les besoins du français. Il convient aussi de prendre conscience de l'importance croissante du multilinguisme, afin de ne pas se cantonner dans des solutions qui ne favoriseraient que le français, solutions myopes qui négligeraient les besoins des langues partenaires et de surcroît priveraient la Francophonie des alliés objectifs qu'elle peut se concilier parmi les locuteurs de toutes les

[2] **Commission électrotechnique internationale** en français, **International Electrotechnical Commission** en anglais, **Международная Электротехническая Комиссия** en russe.

[1] Il s'agit de donner une définition, une explication en contexte pour un mot ou une expression polysémique.

[2] *World Wide Web Consortium*, dont une des trois institutions hôtes est l'INRIA.

autres langues souffrant de problèmes semblables en technologies de l'information. C'est-à-dire presque toutes, bien que l'isolationnisme soit malheureusement fort répandu.

3.2.2. En finir avec les problèmes fondamentaux

Ce sont les problèmes reliés au codage des caractères, qui nuit encore au traitement du français comme on s'en rend compte chaque jour sur Internet. En gros, on peut dire qu'il s'agit d'exorciser l'ASCII, de l'extraire de la position privilégiée qu'il occupe encore dans beaucoup trop d'aspects des technologies de l'information, au grand bénéfice de l'anglais et au grand dam de pratiquement toutes les autres langues.

La première étape est de favoriser l'avènement du JUC, qu'il apparaisse sous les vocables ISO 10646 4, Unicode, UCS ou UTF-8. Les systèmes, logiciels, protocoles ou standards reprenant cette norme de codage devraient être encouragés, favorisés, et les autres dénigrés et remplacés au plus tôt s'il y a lieu.

Il faut ensuite briser l'inertie des cercles de normalisation Internet pour obtenir l'*internationalisation* des protocoles qui sont encore restreints à l'ASCII, mais qui devraient servir à tous de façon égale. On pense au DNS (nommage des machines), aux URL (adresses de ressources Internet, formulaires), aux adresses de messagerie, en fait à toute la messagerie qui est toujours officiellement limitée à l'ASCII et dépend de MIME pour le reste.

1.3.1. Susciter et imposer une norme de tri en Francophonie

Il est important que les normes culturelles et typographiques de tris soient respectées pour le français et ses langues partenaires. Beaucoup de logiciels ou de systèmes informatiques sont livrés avec des routines de tri insuffisantes qui ne répondent pas aux besoins d'un tri correct en français (et encore moins dans certaines langues partenaires). Or, il faut que l'on puisse trier correctement les chaînes de caractères aussi bien en français (ou ses langues partenaires) qu'en anglais. Car un nom ou un titre mal trié ne constituent pas seulement une gêne, mais peuvent souvent résulter dans une apparente absence de l'objet recherché. Les locuteurs de ces langues se voient donc défavorisés par rapport à ceux qui effectueraient une recherche de mots anglais.

Nous avons déjà mentionné à la section 2.1.1 le projet de norme internationale de classement ISO 14651 4. Or une norme rivale, inspirée d'une norme allemande, est malheureusement à l'étude au sein d'un comité européen de normalisation (CEN/TC 304/GTl). Si ce projet devait être retenu, il ne permettrait pas de classer correctement les caractères de la langue française (le classement des caractères accentués serait particulièrement problématique).

Nous recommandons donc dans ce domaine une forte présence francophone dans ces deux forums (ISO et CEN) afin de s'assurer du respect intégral des pratiques francophones en matière de classement alphabétique et de l'harmonisation des normes dans ce domaine. De plus, le fait d'imposer partout où c'est possible une norme de tri *de jure* adéquate forcerait les fournisseurs à développer des logiciels corrects, dont on peut espérer qu'ils seraient ensuite distribués par défaut même là où il n'y a pas d'obligation

légale, au grand bénéfice du français et de ses langues partenaires. L'adoption par l'AFNOR de la norme en élaboration ISO 14651, ou au moins d'une norme plus limitée telle que la norme canadienne CAN/CSA Z243.4.1 4, serait un heureux pas dans la bonne direction.

1.3.2. Augmenter la présence de textes électroniques en français

Il s'agit ici d'augmenter la quantité de documents électroniques en français, de manière à créer une masse critique d'information utile et pertinente pour les citoyens de la Francophonie, et aussi de rendre les inforoutes plus viables en français. Les gouvernements de la Francophonie devraient au minimum favoriser cet objectif en n'insistant pas sur les droits d'auteur des documents qu'ils produisent, avec les deniers publics.

Afin d'inciter un plus grand nombre de particuliers et de sociétés à utiliser l'Internet, condition *sine qua non* à longue échéance d'une présence de services commerciaux rentables et autosuffisants, les gouvernements et organismes assimilés de la Francophonie devraient rendre disponibles en français le plus grand nombre de documents produits par eux :

- prévisions météorologiques ;
- renseignements fiscaux, douaniers ;
- corpus légal, documents administratifs ;
- statistiques, renseignements commerciaux ;
- indicateurs des transports en commun, chemins de fer, lignes aériennes ;
- annuaires téléphoniques et autres...

Les ouvrages de référence (thésaurus, dictionnaires de langue, de synonymes, encyclopédiques, etc.) en français sous forme électronique sont importants comme outils d'aide à la rédaction (voir aussi la section 3.2.3 ci-dessous). Mais ils sont plus que cela, puisqu'ils véhiculent aussi un contenu culturel et peuvent servir à bien plus qu'à la rédaction. Or la forme électronique a d'indéniables avantages par rapport au support papier traditionnel – facilité de recherche, multimédia, etc. – et il importe que le français bénéficie de ces avantages et dispose d'une large gamme d'ouvrages de référence électroniques.

Qui plus est, il faut éviter que ces ouvrages (notamment les encyclopédies) ne soient que des traductions d'ouvrages étrangers, dont le contenu culturel serait immanquablement inadapté. Par contre, il faut encourager l'apparition de corpus bilingues français/langue locale qui pourront servir de base dans la confection de dictionnaires bilingues, d'outils d'aide à la traduction, etc. Évidemment, le format devrait être standard pour tous les corpus francophones mis en commun (cf. la section 3.2.3, Normaliser le codage des ressources linguistiques plus bas).

Chaque année, un grand nombre de thèses sont rédigées en français, soutenues, publiées en quelques copies et, la plupart du temps, oubliées. Nombre d'entre elles, dignes d'intérêt, font l'objet de republication, généralement en forme abrégée, dans les revues

spécialisées du domaine de connaissance, trop souvent en anglais. C'est donc dire qu'un travail intéressant, original et de qualité – puisque soutenu devant un jury de thèse – est fait en français, mais que c'est une version dérivée en anglais qui obtient une grande diffusion.

Or, les thèses sont le plus souvent de nos jours rédigées sur traitement de texte et donc disponibles sous format électronique. De plus, elles ne présentent généralement pas de valeur marchande, ce qui simplifie d'autant les problèmes de droits d'auteur. Il y a donc peu d'obstacles à une publication en ligne de toutes les thèses soutenues en Francophonie, il suffirait de peu de ressources et d'un peu d'effort et de concertation pour institutionnaliser la chose et assurer sa pérennité.

1.3.3. Créer une encyclopédie francophone

La création d'une encyclopédie francophone sous forme électronique, et peut-être même en ligne, pourrait constituer une grande entreprise fédératrice de la Francophonie, un grand projet donnant substance à la communauté des pays francophones. On peut envisager une telle action sous l'angle de la coopération multilatérale, le concevoir comme une œuvre collective de toute la Francophonie.

La mise à disposition en ligne serait idéale au point de vue de la disponibilité, de la rapidité de mise à jour et du rayonnement, mais pose un problème de viabilité économique. Mais même sous d'autres formes (cédérom, etc.), des mesures devraient être envisagées pour favoriser l'accès à cette encyclopédie par les pays du Sud. Il serait ironique, pour dire le moins, que de grands pans de la Francophonie n'aient pas accès à une encyclopédie bâtie en bonne partie par eux et pour eux !

1.3.4. Normaliser le codage des ressources linguistiques

La nécessité de briques de base linguistiques pour les outils de recherche documentaire et pour d'autres applications à caractère linguistique nous amène à nous pencher sur la normalisation et la standardisation des ressources linguistiques.

Il s'agit d'un axe de réflexion extrêmement important dans la mesure où les choix effectués actuellement dans différentes instances de normalisation et de standardisation conditionneront très directement la capacité des francophones à mettre au point les outils de traitement avancé du français dont ils auront besoin pour se tailler une place dans la société de l'information.

Les ressources linguistiques comprennent l'ensemble des matériaux utilisés pour décrire l'état d'une langue et, notamment, ce qui la rend distincte par rapport à une autre. Les dictionnaires, les grammaires, les corpus écrits et oraux, les thésaurus et autres constituent des ressources linguistiques qui sont de plus en plus disponibles sous forme électronique.

Dans le domaine de l'ingénierie linguistique et documentaire, ces ressources sont primordiales dans la réalisation d'applications telles que la génération automatique de textes, les systèmes d'aide à la traduction, les interfaces en langue naturelle, la synthèse ou la reconnaissance vocale, ou encore le repérage documentaire. Il faut cependant savoir

que ces travaux d'établissement de corpus peuvent être extrêmement coûteux. Ainsi, l'entreprise Texas Instruments a-t-elle dépensé plus de 350 000 $US pour concevoir un cédérom reprenant la prononciation de séries de chiffres par quelque 300 personnes. On comprendra donc qu'il est de plus en plus souhaitable d'entreprendre la mise en œuvre des matériaux langagiers de façon concertée et dans un cadre normalisé. L'abaissement de ces coûts d'élaboration par la collaboration, la mise en commun des fonds disponibles et l'utilisation de normes ou de standards devraient permettre au français et à ses langues partenaires d'abaisser les barrières économiques qui s'opposent au plein épanouissement de ces langues dans le domaine des industries de la langue.

Étant donné la jeunesse relative — devrait-on encore parler d'immaturité ? — de ce secteur des industries langagières et sa grande fragmentation, toute tentative de normalisation s'accompagnera d'une sage circonspection. On peut cependant penser que les instances francophones pourraient favoriser cette normalisation en mettant sur pied un fonds des données linguistiques (FODOLIN) auquel contribueraient toutes les universités francophones. Toutes ces universités, et indirectement toutes les entreprises en association avec ces universités, pourraient y bénéficier du résultat des recherches des autres.

Il est souhaitable que l'établissement du FODOLIN ne cherche pas à créer d'énormes structures réparties à travers la Francophonie qui seraient appelées à collaborer sur des projets communs mais bien, dans la grande majorité des cas, des équipes soudées situées en un endroit, dont les membres pourraient provenir de toute la Francophonie. Le FODOLIN ne serait qu'une structure lâche et un lieu de partage avec un strict minimum d'administration, ne cherchant pas à concurrencer les efforts déjà entrepris dans des domaines connexes, mais à les compléter.

Parallèlement à la mise en place de ce fonds, il faudrait poursuivre le travail de normalisation dans ce secteur des industries de la langue. Dans un premier temps, on dresserait un inventaire précis des pratiques existantes en la matière tant en Francophonie européenne que nord-américaine. Cet inventaire devra alimenter la réflexion conduisant à l'établissement de directives et de normes de dépôt (formats permis) au FODOLIN. Le projet européen EAGLES, projet de la DG XIII, qui vise à l'harmonisation des méthodes de création, de description et de représentation des données linguistiques (corpus, lexiques, formalismes) semble avoir porté des fruits, et pourrait peut-être servir de base à une proposition de standardisation.

À court terme, ces directives ne pourront être trop restrictives et seront loin de définir strictement les formats exacts des fichiers mis en commun. On pourra cependant imposer que ces fichiers soient documentés selon les directives prescrites, puis, petit à petit, au fur et à mesure que ce secteur établira des pratiques communes et des standards, on pourra s'en inspirer pour étendre la portée des directives et imposer de la sorte une plus grande normalisation de ce fonds. L'incitation au respect de ces normes pourrait provenir de directives lors de l'octroi de subventions relatives au FODOLIN ou dans des domaines connexes.

1.3.5. Encourager le repérage d'information d'avant-garde

Nous avons déjà discuté de l'évolution du repérage d'information vers la recherche en langue naturelle, évolution prévisible et faisant l'objet de beaucoup de travaux et déjà de certaines réalisations d'avant-garde. Qu'il s'agisse des inforoutes ou de grandes bases de données plus localisées (par ex. intranet), il est clair que ce type de repérage peut être très bénéfique pour les utilisateurs ; on assistera donc à une vive croissance si les réalisations (y compris en langue française) tiennent la route. Il est donc impératif d'inciter l'apparition d'outils de repérage (pourrait-on les appeler « glaneurs » ?) dotés d'une très bonne compréhension du français et de ses langues partenaires. Dans ce sens, il est à nouveau préférable de favoriser le développement d'outils multilingues. Il faut encourager, au sein des centres de recherche et des universités, le développement d'une expertise dans le domaine de l'analyse morpho-syntaxique, de la désambiguïsation, de thésaurus, de glosage en français et d'autres techniques linguistiques pour chacune des grandes langues de la francophonie. Ces projets sont également réalisables dans une certaine mesure[1] dans le Sud. Ces éléments pourraient servir comme bases d'autres projets : aides à la rédaction, à la traduction ou même la traduction automatique.

Mais même si l'on pouvait s'adresser en langue naturelle à ces glaneurs, on peut craindre que les requêtes ne doivent être formulées avec beaucoup de soin pour atteindre les objectifs habituels de rappel et de précision. Pour améliorer le rappel, une voie d'avenir que nous suggérons ici serait de développer des méthodes de repérage en langue naturelle basées sur les figures de rhétorique classiques comme la métonymie[2], la métaphore[3] ou l'analogie[4], qui permettraient d'élargir le champ de la recherche, de « ratisser plus large ».

1.3.6. Estimer la fiabilité de la documentation en ligne

L'explosion récente des inforoutes a eu l'effet très bénéfique de mettre à disposition une grande quantité d'information en ligne, mais l'exploitation de cette bibliothèque géante et délocalisée ne manque pas de poser certains problèmes. Celui de la recherche d'informations pertinentes est bien connu et fait l'objet de recherche et de

français constituant une des richesses de la Francophonie. De tels projets existent d'ailleurs, par exemple l'Action de Recherche Partagée intitulée « Acquisition automatique de terminologies dans des corpus de langues africaines et françaises » qui s'est penché sur le français, l'anglais, le zarma (Niger) et le malgazy (Madagascar). Cf. La lettre d'information Francil numéro 9, octobre 1997, pp2-3.

[2] **métonymie** *nom féminin* (grec *metônumia,* changement de nom)
Ling., rhét. Phénomène par lequel un concept est désigné par un terme désignant un autre concept qui lui est relié par une relation nécessaire (l'effet par la cause, le contenu par le contenant, le tout par la partie, etc.). [Ex. : *il s'est fait* refroidir (tuer) ; *toute la* ville *dort* (les habitants) ; *une fine* lame (escrimeur) ; etc.] © Larousse.

[3] **métaphore** *nom féminin* (grec *metaphora,* transport)
Ling., rhét. Procédé par lequel on transporte la signification propre d'un mot à une autre signification qui ne lui convient qu'en vertu d'une comparaison sous-entendue. (Ex. : *la* lumière *de l'esprit, la* fleur *des ans,* brûler *de désir, ficelle* au sens de « pain », etc.) © Larousse.

[4] **analogie** *nom féminin* (grec *analogia*)
1. Rapport de ressemblance que présentent deux ou plusieurs choses ou personnes. *Analogie de forme, de goûts. – Par analogie :* d'après les rapports de ressemblance constatés entre deux choses.
2. Ling. Apparition dans une langue de nouvelles formes à partir de correspondances qui existent entre des termes d'une même classe. © Larousse.

développement tous azimuts. Mais il en est un autre qui est, sinon inconnu, du moins un peu négligé, et peut donc constituer une occasion intéressante pour la Francophonie : l'évaluation de la qualité, de la pertinence et de la fiabilité des ressources trouvées en ligne.

La solution traditionnelle (hors inforoutes) à ce problème est basée sur un heureux mélange de bibliothécaires, de jugement par les pairs dans les publications scientifiques, de réputation et de spécialisation des éditeurs ou des auteurs connus, etc. Mais bien peu de ce système a sa contrepartie sur les inforoutes, où n'importe qui peut publier n'importe quoi n'importe quand, et ce bien souvent d'une adresse où le bon grain et l'ivraie sont intimement mélangés.

On peut envisager l'établissement d'un (ou de plusieurs) système de référence pour résoudre ce problème ; le système PICS[1] peut être considéré comme une ébauche timide d'un tel système, mais on peut imaginer beaucoup mieux.

D'une part, une analyse linguistique du texte peut permettre de juger du style et de la correction grammaticale, donnant une idée d'au moins un aspect de la qualité étant entendu que l'auteur d'un document de qualité prendra quelque soin de sa forme. Une analyse plus poussée, incluant un volet sémantique, peut permettre la vérification de faits, d'affirmations faites dans le document ; la confirmation d'un fait augmente alors la crédibilité du document, alors qu'une erreur reconnue la diminue.

D'autre part, on peut ajouter au système un réseau de fiabilité, un peu comme dans les systèmes de chiffrement à clés publiques ou, plus pertinemment comme le *Science Citation Index* et *Social Science Citation Index* (en anglais seulement) utilisés par la communauté scientifique. Si le document qui m'intéresse est cité par Untel, et si j'ai confiance en Untel, alors le document est plus crédible. Or ma confiance en Untel peut provenir d'autres citations, et non pas d'une connaissance directe d'Untel, d'où la notion de réseau. La même chose fonctionne bien sûr en sens inverse (exclusions, liste noire), d'où en fait la nécessité d'une analyse des citations : est-ce qu'Untel attaque le document qui m'intéresse ou le loue ? Ce réseau de fiabilité peut sembler moins pertinent au traitement informatique du français, mais en réalité ce dernier ne se produit jamais en vase clos, il fait toujours partie d'une application visant à une certaine utilité.

1.3.7. Favoriser la création et la dissémination d'outils d'aide à la rédaction

Il s'agit ici de rendre la rédaction en français au moins aussi facile que la rédaction en anglais. Des outils puissants et conviviaux (vérificateurs orthographiques, grammaticaux, conjugueurs, etc.) facilitant la rédaction en français peuvent contribuer à rendre cette langue plus attrayante pour un rédacteur cherchant une langue véhiculaire. Ces outils peuvent aussi susciter une meilleure qualité du résultat, permettant par la suite une exploitation – par d'autres outils d'ingénierie linguistique – plus aisée, plus précise et/ou plus fertile.

[1] *Platform for Internet Content Selection*. Ce système est principalement destiné à protéger la jeunesse des sites inconvenants. Cf. http://www.w3.org/PICS/.

On ne doit pas oublier non plus les technologies d'extraction automatique d'information (cf. section 2.4.1) dans lesquelles la Francophonie semble prendre un sérieux retard. Il conviendrait d'encourager des concours amicaux francophones du type *MUC*, ou au minimum de participer à celle-ci.

1.3.8. Favoriser la création et la dissémination d'outils d'aide à la traduction

Sachant qu'une très grande part de l'information sous forme électronique est actuellement en anglais, que ce soit sur les inforoutes, dans les intranets ou ailleurs, il importe de faciliter la traduction de cette information en d'autres langues pour minimiser le besoin de recours à l'anglais. On vise ainsi à pallier la marginalisation du français qui se produira si on laisse l'anglais devenir la seule langue dans laquelle il est possible d'obtenir une information complète. Il est donc essentiel de faciliter la traduction **vers** le français (et ses langues partenaires, en même situation), d'améliorer la vitesse de traduction et d'en réduire les coûts. Lorsqu'on sait que, par exemple, l'Office européen des brevets cherche par sa « solution globale »[1] à supprimer les traductions à cause de leur coût, cet intérêt ne fait aucun doute.

Il importe donc que les outils informatiques d'aide à la traduction, allant du traitement de texte aux mémoires de traduction, en passant par les dictionnaires (électroniques) et les bases de terminologie, soient à la fine pointe et bien harmonisés pour atteindre ce but. Une voie d'avenir fort prometteuse se manifeste aussi dans la traduction interactive 4 4, qu'on pourrait aussi qualifier de semi-automatique, en fait une forme de traduction assistée par ordinateur (TAO) : la machine tente de traduire le texte, mais peut poser des questions à un opérateur – peut-être l'auteur ou un relecteur – pour résoudre les problèmes qu'elle rencontre et pour lui demander de lever les ambiguïtés. Une autre approche légèrement différente 4 4, mais visant au même résultat, consiste à programmer la machine prend en compte deux sources d'information, à savoir le texte en langue de départ et une traduction partielle de ce texte produite par le traducteur humain pour proposer une traduction plus complète. Dans cette nouvelle approche, c'est toujours le traducteur qui mène. La traduction automatique est utilisée comme une aide à la frappe. On peut en espérer une qualité de traduction bien supérieure à la TA pure et simple, en même temps qu'une vitesse, et donc un coût de revient, bien meilleure qu'en traduction humaine.

1.3.9. Favoriser le rayonnement du français par la TA ou la TAO vers d'autres langues

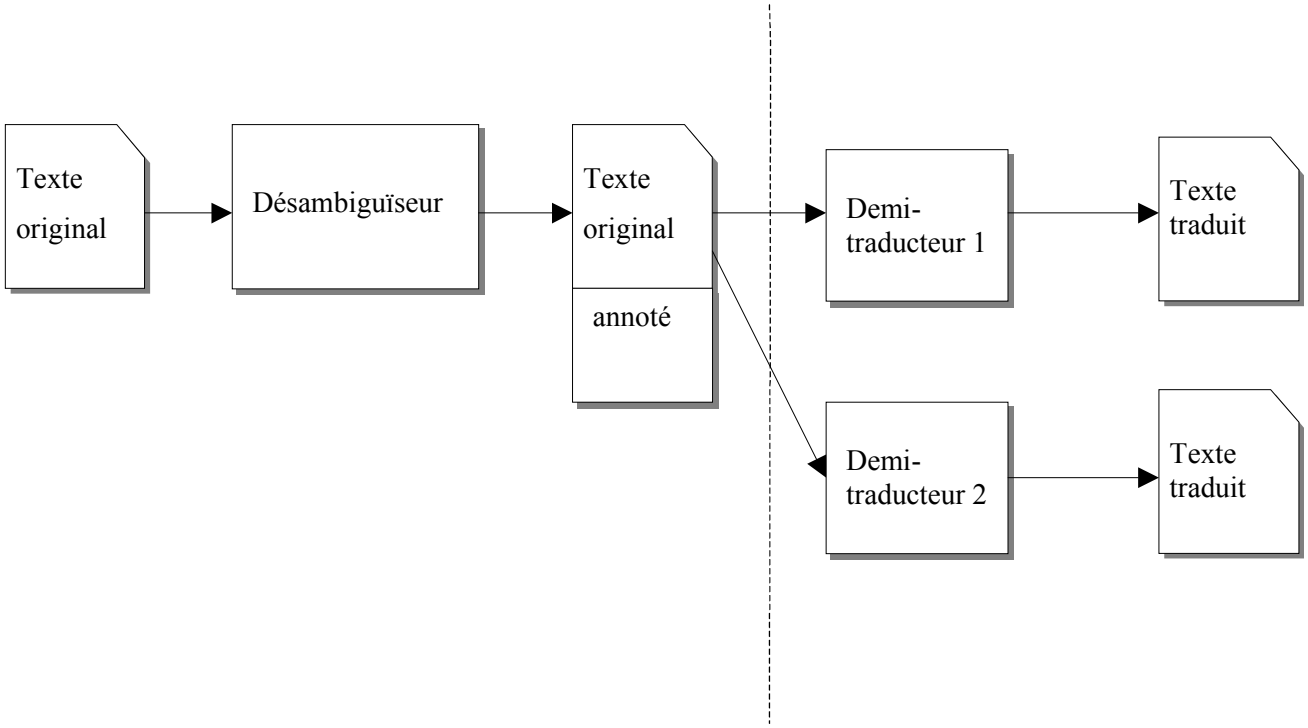
La marginalisation menace aussi le français si on ne parvient pas à traduire efficacement **du** français vers d'autres langues. Par exemple, on commence à voir des applications de la traduction automatique dans le domaine de la publication en ligne[2] : les documents rédigés en une langue source sont publiés en plusieurs langues par le biais de la TA. Là où

[1] L'O.E.B. a proposé à la Commission européenne un livre vert, *Promouvoir l'innovation par le brevet :Livre Vert sur le brevet communautaire et le système des brevets en Europe*, qui contient une proposition de « solution globale » éliminant pratiquement les traductions de brevets européens, en contradiction avec plusieurs lois nationales. Cf. http://europa.eu.int/comm/dg15/fr/intprop/indprop/558.htm, mais surtout le livre vert lui-même, accessible depuis cette page.

[2] C'est une des applications du produit Solutions de traduction d'Alis. Cf. http://www.alis.com/ats/index.fr.html.

le bât blesse, c'est qu'il n'est présentement possible de ce faire que si les originaux sont rédigés en anglais, les moteurs de traduction du français vers d'autres langues n'étant pas disponibles ou de qualité insuffisante[1]. L'éditeur doit donc se passer de cette méthode de publication multilingue à bon marché, ou encore engager des rédacteurs anglophones, au détriment bien sûr de l'emploi de francophones.

Aussi bien la TA que la TAO peuvent être bénéfiques ici, selon les applications, et il convient de les encourager. Dans un mode plus spéculatif, on peut aussi envisager un modèle particulier de traduction interactive à retardement, où la première étape d'analyse du texte source, avec questions-réponses, par un *désambiguïseur* est séparée de l'étape subséquente de génération du texte traduit. Le résultat de la première étape peut être considéré comme un texte *annoté* libéré de toute ambiguïté, et donc éminemment traduisible en n'importe quelle langue[2] par la seconde moitié d'un moteur de TA adapté (un *demi-traducteur*). Si de tels demi-traducteurs existaient pour plusieurs langues, la publication du texte annoté aurait donc beaucoup plus de valeur que la publication du même texte dans une autre langue qui ne bénéficierait pas d'un tel mécanisme.



Sans aller jusqu'à la traduction, le glosage déjà mentionné peut aussi aider au rayonnement du français, en permettant à ceux qui le connaissent encore mal de lire des

[1] La qualité des moteurs de traduction de l'anglais n'étant que très marginalement suffisante pour ce genre d'application, le moindre écart vers le bas nous amène dans la catégorie « insuffisant ».

[2] À dire vrai, cette affirmation doit être tempérée. Par exemple, bien que la phrase française « Je vais mettre de l'eau à chauffer » semble exempte de toute ambiguïté, sa traduction en japonais exige de faire le choix entre les mots *yu*, « eau chaude » et *mizu* « eau » : *o* **yu** *o wakashite kimasu* dans ce cas, bien que l'eau ne soit pas encore chaude ! Il est peu probable que le texte annoté tienne compte de cette distinction, à moins que de *désambiguïseur* n'ait été conçu en fonction d'une traduction vers le japonais.

textes et d'obtenir au besoin des définitions correctes *en contexte* de termes qu'ils ne comprendraient pas. Le glosage peut donc être considéré comme une aide à la lecture, complétant la gamme des aides à la rédaction (outils ou ouvrages de référence) et des aides à la traduction.

1.3.10. Mettre en ligne des services linguistiques

Les organismes tels que l'OLF[1] devraient mettre en ligne leurs services linguistiques traditionnels. Des frais pourraient être perçus pour certaines prestations longues et coûteuses. On peut imaginer une vaste gamme de services :

- terminologie ;
- néologismes ;
- traduction ;
- grammaire ;
- enquête en ligne ;
- bulletins d'information de l'actualité langagière.

[1] Office de la langue française du Québec.

4. Références

[1] **Palmarès des langues de la Toile**, Alis Technologies, juin 1997, http://babel.alis.com:8080/palmares.fr.html.

[2] Gaston Berger, *Éducation et enseignement dans un monde en accélération,* in **L'homme moderne et son éducation**, Paris, P.U.F., 1967, pp118-199.

[3] ISO/CÉI 10646-1 – **Technologies de l'information — Jeu universel de caractères codés sur plusieurs octets (JUC) — Partie 1 : architecture et plan multilingue de base**.

[4] ISO/CÉI 9995 : 1994 – **Technologies de l'information – Disposition des claviers conçus pour la bureautique**, 8 parties, dont une seule publiée en français, fruit du travail d'un Québécois.

[5] CAN/CSA Z243.4.1 – **Méthode canadienne de classement applicable aux normes CAN/CSA-Z243.4 et CSA T500**.

[6] ISO/CÉI CD 14651 – **Classement international de chaînes de caractères – Méthode de comparaison de chaînes de caractères et description d'un ordre de classement implicite adaptable** – (en préparation).

[7] RFC 1869, **SMTP Service Extensions**, J. Klensin, N. Freed, M. Rose, E. Stefferud & D. Crocker, Novembre 1995.

[8] RFC 2045, **Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies**, N. Freed & N. Borenstein, Novembre 1996.

[9] RFC 2046, **Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types**, N. Freed & N. Borenstein, Novembre 1996.

[10] RFC 2047, **MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text**, K. Moore, Novembre 1996.

[11] RFC 2048, **Multipurpose Internet Mail Extension (MIME) Part Four: Registration Procedures**, N. Freed, J. Klensin & J. Postel. Novembre 1996.

[12] RFC 2049, **Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples**, N. Freed & N. Borenstein, Novembre 1996.

[13] Kukich, K. **Knowledge-Based Report Generation: A Knowledge-Engineering Approach**. thèse de doctorat, Université de Pittsburgh, 1983.

[14] Goldberg E., Dreidger N. and Kittredge R., *FoG: A New Approach to the Synthesis of Weather Forecast Text*, IEEE-Expert, 9:2, 1994, pp45-53.

[15] Darragh J.J. & Witten I.H. **The Reactive Keyboard**, Cambridge University Press, 1992.

[16] Pasero R., Richardet N. , Sabatier P., *Guided Sentences Composition for Disabled People*, Proceedings of ANLP-94, Stuttgart, 1994.

[17] Bar-Hillel, Y. *The State of Machine Translation in 1951*, **American Documentation**, vol. 2, 1951, pp229-237.

[18] Chevalier M., Dansereau J. & Poulin G., **TAUM-MÉTÉO: description du système**, publication interne du groupe TAUM, Université de Montréal, 1978.

[19] **Trial of the Weidner Computer-Assisted Translation System**, Project No. 5-5462, Bureau of Management Consulting, Department of Supply and Services, Govt. of Canada, octobre 1985.

[20] Kay, Martin, *The Proper Place of Men and Machines in Language Translation*, Tech. Rep. no. CSL-80-11, Xerox PARC, 1980. À paraître dans **Machine Translation**.

[21] **Applications of natural language understanding**, Gartner Research Note, 23/08/95, Advanced Technologies & Applications.

[22] John Naisbitt, **Global Paradox**, William Morrow and Company, New-York 1994, 304 pp. (ISBN 0-688-12791-6). La citation est en page 26.

[23] Éric Wehrli, *Vers un système de traduction interactif*, in P. Bouillon et A. Clas (éd.), **La Traductique**, Les Presses de l'Université de Montréal, 1993, pp423-432.

[24] Éric Wehrli, *Traduction interactive : problèmes et solutions (?)*, in A. Clas et P. Bouillon (éd.), **TA-TAO : Recherches de pointe et applications immédiates**, Montréal, Aupelf-Uref, 1994, pp333-ss.

[25] Foster G., Isabelle P., Plamondon P., *Word Completion: a First Step Toward Target-Text Mediated MT*, Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), Copenhague, 1996.

[26] Foster G., Isabelle P., Plamondon P., *Interactive Machine Translation by Target Text Mediation,* **Machine Translation**, 12:1-2, 1997, pp175-194.