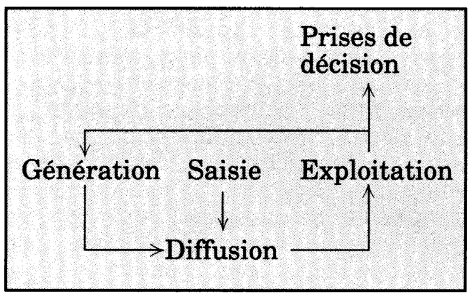
Les produits « infolinguistiques » : un tour d'horizon

Dans les sociétés postindustrielles, le mot est en voie de détrôner le chiffre comme véhicule principal d'information; à l'évidence, la maîtrise des langues naturelles devient un enjeu majeur de concurrence dans le monde de demain.

C'est ce qui explique la fébrilité avec laquelle le secteur des industries de la langue se développe depuis quelques années à peine. Ces industries travaillent à des produits que l'on peut appeler « infolinguistiques ». Il s'agit de produits informatiques qui permettent ou facilitent les opérations portant sur l'information textuelle : sa création, sa saisie, sa diffusion, sa transformation et son exploitation. À ces opérations s'ajoute aussi la communication en langage naturel entre humains et machines.

Le tour d'horizon proposé ici s'articule non pas sur la technologie des produits infolinguistiques, mais sur leur fonction. Il vise à expliquer non pas de quoi ils sont faits, mais à quoi ils servent. Cette approche est liée à une définition élargie des industries de la langue, qui englobe l'ensemble des activités langagières assistées par ordinateur. Ces activités s'inscrivent dans l'ensemble du *cycle de l'information*, que l'on pourrait décrire par la boucle suivante :



Le sujet de la communication humain-machine ne fait pas partie de ce schéma, et sera traité à part.

L'étape de la génération

Le traitement de texte, désormais banalisé, constitue le nerf de la génération de texte. Depuis quelques années, on voit s'y greffer toutes sortes de fonctions auxiliaires.

Il y a bien sûr l'*éditique* (ou *micro-édition*, ou *publication assistée par ordinateur* [PAO]), qui concerne la mise en page définitive de toute la gamme des éléments textuels et graphiques nécessaires à l'obtention d'un original prêt à imprimer ou à reprographier. L'éditique accélère ce processus; en outre, étant donné son prix de plus en plus abordable, elle tend à le rendre accessible à tous.

Disons-le, la frontière entre l'éditique et le traitement de texte devient floue à mesure que ce dernier offre des fonctions de plus en plus perfectionnées (polices typographiques, formatage avancé, feuilles de style...). Dans les années 90, nous assisterons sans doute à l'intégration complète de ces deux technologies.

Les « critiqueurs[1] » de texte concernent l'acte même d'écrire. Leur grand avantage est de repérer les difficultés non soupçonnées : ils jouent en quelque sorte le rôle inestimable de « chien de garde » ou d'aide-mémoire. Le critiqueur *orthographique*[2] est de loin l'outil le plus courant : plusieurs grands traitements de texte (Word, WordPerfect, XY-Write, etc.) en offrent un en français; outre ces correcteurs internes, d'autres correcteurs sont vendus séparément et fonctionnent de façon externe au traitement de texte. Signalons aussi que l'emploi d'un critiqueur orthographique ne dispense pas d'une relecture attentive : si une faute correspond accidentellement à un mot correct (exemple : « ont ta prit thon jouais... »), le critiqueur n'y voit que du feu.

Les critiqueurs *grammaticaux* sont assez rares en français; le seul produit de cette catégorie, HUGO (diffusé par Logidisque), semble de qualité douteuse. En anglais, la question se pose différemment; étant donné le peu de flexions de cette langue, les critiqueurs dits « grammaticaux » (Correct Grammar, Grammatik III, Right Writer, MacProof, etc.) ont peu de chose à se mettre sous la dent au chapitre des accords, et englobent invariablement l'aspect stylistique (voir ci-après). Le plus avancé de ces produits est certainement Critique d'IBM, réservé pour le moment aux gros ordinateurs.

Enfin, les critiqueurs *stylistiques* détectent les fautes d'usage (solécismes, barbarismes, etc.) et permettent d'améliorer divers paramètres stylistiques (longueur des phrases, niveau de langue, sexisme, charabia, etc.). Signalons leur potentiel à servir d'encadreurs pour les rédacteurs de documentations spécialisées. C'est ainsi que la société américaine Smart Communications Inc. commercialise depuis plusieurs années un critiqueur de texte technique anglais appelé SMART Expert Editor; ce genre d'outil, qui sera certainement de plus en plus employé dans l'industrie de la rédaction, aide à normaliser l'écriture entre plusieurs rédacteurs, filtre les éléments pouvant donner lieu à des poursuites, et peut aussi rendre le texte plus facile à traduire par ordinateur. Cela dit, il n'existe, semble-t-il, aucun critiqueur stylistique commercial pour le français.

Il existe maintenant une solution de rechange à la rédaction au clavier : la rédaction vocale, au moyen de la légendaire *machine à dicter automatique* (que certains appellent assez joliment *vocoscribe*). Dans ce domaine, les rares produits commerciaux proviennent des États-Unis; les deux grands développeurs sont probablement Kurtzweil Computer Products et Dragon Systems. La technologie existe, mais ses performances encore chancelantes obligent à divers compromis : tantôt la machine ne comprend que la voix d'une seule personne (mode monolocuteur), tantôt il faut lui parler à mots détachés, tantôt encore on ne peut employer qu'un vocabulaire limité. Sans parler du taux d'erreurs, assez variable. Pour que l'utilisation de telles machines puisse se généraliser, ces lacunes devront être comblées et, détail important, le prix d'achat ne devra pas être trop élevé en regard des avantages retirés.

À part la fascination qu'exerce la technologie vocale, on peut donc douter d'une invasion prochaine du vocoscribe; son marché privilégié, pour l'instant, est surtout celui des cadres « ordinophobes ». Entre-temps, le clavier — que la pratique de la micro-informatique contribue à implanter dans les bureaux — restera longtemps concurrentiel pour la masse des travailleurs de bureau.

Après les activités de rédaction, la *traduction* a droit à sa panoplie de produits infolinguistiques. Les plus courants sont des outils de consultation de lexiques bilingues, comme Termex et TermTracer, qu'on peut employer à même la fonction traitement de texte.

De plus en plus, on reconnaît aussi que de bons outils bureautiques de base peuvent augmenter à peu de frais la productivité du traducteur. C'est ainsi que le Centre canadien de recherche sur l'informatisation du travail (CCRIT) a mis au point sur PC un prototype appelé PTT (poste de travail du traducteur), regroupant essentiellement divers logiciels commerciaux à l'intérieur d'un environnement multitâche. Un traducteur d'Ottawa, Michel Thibodeau, a fait de même pour le Macintosh[3].

Les systèmes de *traduction automatique* sont maintenant nombreux; citons Systran, Logos, MicroCAT, TOVNA et METAL pour les langues européennes. De plus en plus, grâce notamment aux progrès de la micro-informatique, de nouveaux systèmes sont offerts à des prix abordables et tournent sur micro-ordinateur; deux exemples sont PC-Translator (Linguistic Products) et GTS (Globalink), tous deux développés aux États-Unis. À part de rares exceptions notables comme MÉTÉO, les systèmes de traduction automatique continuent d'offrir des performances imparfaites; par contre, il faut dire que les utilisateurs n'ont pas non plus acquis tout le savoir-faire souhaitable pour tirer le meilleur parti possible de ces systèmes.

L'étape de la diffusion

Une information qui ne circule pas ne se développe pas. Qu'à cela ne tienne! L'informatique moderne ouvre à la diffusion de l'information textuelle des perspectives très intéressantes.

La *télématique* — qui, disons-le, piétine quelque peu depuis les années 70 — devrait se développer de façon décisive dans les années 90. Sur le plan technologique, la récente norme internationale RNIS (réseau numérique à intégration de services) permettra la numérisation complète du réseau téléphonique d'ici quelques années, offrant ainsi un formidable potentiel de maillage informatique. Les banques de données publiques sont maintenant nombreuses; au Québec, la principale société dans le domaine est sans doute Services documentaires Multimédia (un de ses produits les plus connus est Logibase, une base de données sur les logiciels québécois). Le lancement du service ALEX de Bell (concurrent canadien du Minitel) encouragera le développement — et l'utilisation — de banques de données.

Mais c'est certainement l'imprononçable *disque optique compact*[4] (CD-ROM) qui excite l'imagination. D'une capacité dépassant (pour le moment) les 600 méga-octets, il fait ressortir instantanément certains désavantages du support papier (coût de production et difficulté de consultation) et de la télématique (tracasseries à la connexion, frais de consultation à la minute, etc.). On peut s'attendre à une progression rapide de cette technologie maintenant que les grands de l'électronique mondiale se sont entendus à son sujet sur la norme ISO 9660. En fait, le disque optique concurrence d'ores et déjà le livre dans le secteur des ouvrages de référence. Citons très sommairement Bookshelf (Microsoft), qui regroupe plusieurs dictionnaires et autres ouvrages de consultation courants, CHOIX (Services documentaires Multimédia), qui donne 230 000 titres de livres et publications en langue française, l'encyclopédie américaine *Grolier*, le *Grand Robert électronique*. Il est clair que les grands éditeurs voient dans le disque optique un support de publication auquel ils doivent logiquement s'intéresser de près. Gageons que dans cinq ans toute librairie qui se respecte aura son rayon (sous clé) d'ouvrages sur disque optique.

Rappelons quand même que si le disque optique contourne allégrement la télématique pour les données peu périssables, celle-ci conserve quand même l'avantage (déterminant dans bien des cas) d'une mise à jour constante.

L'étape de l'exploitation

Pour bien fermer la boucle, il ne nous reste plus qu'à prendre connaissance de cette montagne d'informations, de plus en plus pléthorique et redondante et, pour notre plus grand bonheur, de mieux en mieux diffusée... soit pour éclairer nos prises de décision, soit encore dans le but de contribuer à notre tour à ladite montagne.

Heureusement, nous pouvons compter maintenant sur une *informatique documentaire*[5] de plus en plus performante. On peut distinguer deux grandes étapes : d'abord, retrouver dans la masse des documents ceux qui sont pertinents (l'étape extra-textuelle); ensuite, à l'intérieur d'un texte donné, repérer exactement les passages pertinents (l'étape intra-textuelle).

La première étape fait intervenir les *gestionnaires de bases de données* ou systèmes de gestion de base de données (SGBD). Ces systèmes procédaient classiquement par la technique bien connue de mots clés choisis manuellement; de plus en plus, on voit se développer des SGBD plein texte, qui tiennent compte de tous les mots du texte (sauf les mots-outils comme les articles, les prépositions, etc.). Signalons à titre d'exemples deux SGBD plein texte : Édibase (Inform II-Microfor) et Seconde (Destin inc.).

Deuxième étape : l'analyse de texte, ou lecture orientée . Un outil particulièrement intéressant est SATO, logiciel développé par le Centre d'ATO (analyse de textes par ordinateur) de l'Université du Québec à Montréal. Son trait original est la capacité d'annoter le texte, c'est-à-dire d'adjoindre aux mots des caractéristiques qui pourront être utilisées pour l'interrogation; les résultats de l'interrogation peuvent eux-mêmes servir à une phase d'annotation, et ainsi de suite. Un outil comme SATO permet véritablement d'ajouter de la valeur au texte. Une de ses applications favorites consiste à projeter sur le texte le lexique de termes du domaine, de façon que ceux-ci soient traités comme tels et non comme la somme des mots qu'ils contiennent.

... Et la saisie, alors!

Les descriptions qui précèdent montrent bien l'importance et l'utilité des traitements informatiques qu'on peut effectuer sur l'information textuelle... à condition qu'elle existe préalablement sous forme ordinolingue (ou comme on dit plus couramment, sur support informatique). D'où l'intérêt des moyens de *saisie* — qu'on a trop souvent tendance à passer sous silence et qui pourtant constituent une porte d'entrée vitale au « tout-informatique » textuel[6].

Comme le support universel de l'information textuelle est le papier, la *lecture optique de caractères* (LOC) est un médium très sollicité de nos jours. Les systèmes de lecture optique se présentent généralement sous la forme d'un lecteur optique (qui opère la saisie d'images), d'un logiciel de reconnaissance de caractères et d'un micro-ordinateur. En haut de la gamme, on trouve Kurtzweil, le pionnier de la lecture optique (dont la version PC se vend maintenant à un prix à peu près abordable); en bas, les produits sont maintenant innombrables, et l'on peut citer ReadStar (Inovatic S.A.) et PC Scan (DEST). Les performances varient largement et dépendent de toutes sortes de variables : qualité de l'original, mise en page, police de caractères, etc. En deçà d'un taux de succès très élevé, l'opération n'est pas tellement profitable (qu'on y pense : un taux de 98 % signifie en gros deux fautes pour cent caractères, soit plus d'une faute par ligne). La détection des erreurs, bien sûr, gagne à employer un critiqueur orthographique. Encore là, le savoir-faire de l'utilisateur fait une certaine différence.

Enfin, on peut assimiler à la saisie les outils qui permettent à un matériel informatique d'accueillir l'information d'un autre, en principe incompatible. Les différents logiciels de *conversion* nous sont de plus en plus familiers : KeyWord (KeyWord Office Technologies) pour la gamme des traitements de texte dédiés; Software Bridge (Systems Compatibility Corp.) pour les

traitements de texte sur PC; et enfin MacLink Plus (DataViz Inc.) pour les conversions entre PC et Macintosh. Ces outils, pour humbles qu'ils soient, sont une véritable bénédiction pour leurs utilisateurs et méritent d'être mentionnés.

Les interfaces en langage naturel

Outre les opérations qui portent sur l'information elle-même, les industries de la langue travaillent à réaliser une fonction encore insuffisamment développée : le dialogue entre l'humain et l'ordinateur. À mesure que l'informatique se généralise, les utilisateurs trouvent de moins en moins le temps de maîtriser le code des commandes attachées à la logique de fonctionnement de chacun des logiciels. La solution : une interface dite « en langage naturel » (ILN) permettant à l'utilisateur d'exprimer son besoin dans ses propres mots et, le cas échéant, de comprendre en retour les réponses de la machine.

Dans une interface en langage naturel, le traitement informatique consiste d'une part à *analyser* les messages de l'utilisateur dans le but de les comprendre, et d'autre part à *générer* des réponses en langage naturel. Le dialogue est soit *écrit* (l'humain tape au clavier et les répliques de l'ordinateur s'affichent à l'écran), soit *oral* (l'humain parle dans un micro et l'ordinateur lui répond de sa voix synthétique). Que le dialogue soit écrit ou oral, trois étapes sont nécessaires. D'abord, il faut que la machine fasse une analyse linguistique de ce qu'on lui dit, puis qu'elle procède à une compréhension conceptuelle, et enfin qu'elle décide de la façon de réaliser la chose demandée.

Actuellement, l'application la plus répandue est l'interrogation des bases de données; citons Intellect (Artificial Intelligence Corporation), qui ne fonctionne que sur gros ordinateur, et aussi Clout (Microrim), NaturalLink (Texas Instruments) et Savvy (Excalibur Technologies) pour micro-ordinateur. Deux autres applications notoires sont l'interrogation des *systèmes experts* (SE) et l'*enseignement intelligemment assisté par ordinateur* (EIAO). Dans l'avenir, on peut imaginer que n'importe quel système informatique disposera d'un « pilote automatique » qui permettra à l'utilisateur, à partir de commandes en langage naturel, d'obtenir l'effet voulu sans avoir à chercher interminablement dans le manuel de référence.

Notons encore ici que la composante *vocale* dans une interface en langage naturel, certes fascinante, n'ajoute pas une dimension essentielle; elle permet de lâcher le clavier, mais sa ressource principale est davantage le travail *cognitif* qu'elle accomplit. Bien qu'à terme l'interface vocale soit appelée à se généraliser, elle n'est pour l'instant indispensable que dans certaines situations souvent sans réel dialogue : commande de machines par des personnes handicapées, tâches occupant déjà totalement les mains ou les yeux (chirurgie, contrôles de qualité, avions de combat), accès par réseau téléphonique (*télématique vocale*) si l'usager ne dispose pas d'un terminal avec modem, etc.

La composante linguistique

Vous aurez peut-être remarqué que la composante linguistique est présente de façon très variable dans les produits infolinguistiques. Certains produits, comme le traitement de texte, les logiciels de conversion et les disques optiques numériques, ne font guère appel à des techniques dites de « traitement des langues naturelles » (TLN). Si l'on peut affirmer sans risque que la composante linguistique finira par infiltrer à peu près tous les recoins des technologies liées à l'information, il est certain aussi que les industries de la langue font appel à toutes sortes de fonctionnalités qui n'ont pas grand-chose de linguistique — il faut bien vivre! D'où l'intérêt justement de définir les industries de la langue par ce qu'elles permettent de faire plutôt que par ce qu'elles sont intrinsèquement.

Cet ensemble de fonctionnalités souvent assez générales (linguistiques ou non), chaque discipline en – *tique* y puise largement, selon ses besoins propres. Pour la *terminotique*, par exemple, on peut citer la télématique, l'informatique documentaire, les SGBD, la saisie optique, l'analyse de textes, l'éditique, le disque optique... Le seul outil vraiment spécifique serait un analyseur de textes axé sur le repérage des termes et sur la récolte de leurs descripteurs; c'est l'objet du projet Termino (Office de la langue française / Centre d'analyse de textes par ordinateur), actuellement en cours.

Développer c'est bien, utiliser c'est mieux

Pour conclure ce tour d'horizon axé sur le rôle des produits infolinguistiques, il est important de faire la part des choses entre la création de ces produits et leur utilisation.

Quand on parle d'industries de la langue, il est à peu près toujours question de créer des produits et de les vendre; on se place volontiers du point de vue du développeur, et l'on s'intéresse alors au potentiel commercial, au chiffre d'affaires. Les industries de la langue sont vues comme un secteur fournisseur de produits infolinguistiques. Par contre, on pense moins à mettre l'accent sur l'*utilisation* de ces produits, qui permet pourtant des augmentations de productivité dont le poids économique est incomparablement plus important; qu'on pense simplement aux gains de productivité réalisés en utilisant un logiciel de traitement de texte, par rapport à son prix d'achat! Le fait d'étendre la notion d'industries de la langue aux utilisateurs permet de rendre justice à cet aspect hautement stratégique du point de vue de l'ensemble de l'économie.

D'ailleurs, l'acceptation et la maîtrise des produits infolinguistiques par l'utilisateur compte certainement pour la moitié du travail à accomplir pour réaliser le plein potentiel des produits infolinguistiques. Or l'inertie, le confort des habitudes acquises pèsent d'un poids considérable. L'utilisateur s'attend trop souvent à travailler *de la même façon* avec les outils plus performants qui lui sont offerts : c'est dans cet esprit qu'il juge de l'utilité des produits, puis qu'il s'en sert (pas toujours bien). On ne le répétera jamais assez : chez les « travailleurs du texte » en particulier, l'imagination et l'audace sont des ingrédients indispensables au succès de la technologie infolinguistique. Le progrès, c'est autant dans la tête que sous les doigts... Et l'utilisateur doit assumer sa part de responsabilité à cet égard.

Claude Bédard

Titulaire d'une maîtrise en traduction de l'Université de Montréal, Claude Bédard est l'auteur de plusieurs ouvrages et articles sur la traduction technique. Il est aussi consultant en traduction assistée par ordinateur et correspondant canadien de la revue Electric Word.

Notes

1. Nous préférons employer *critiqueur* plutôt que *correcteur*, étant donné que le logiciel ne fait que signaler les fautes possibles; c'est l'humain qui décide s'il y a effectivement faute, et qui opère la correction.

2. Voir l'article « Tout sur la détection des fautes d'orthographe », par John Chandioux, dans *L'Actualité terminologique*, vol. 22, n° 3, p. 15-17.

3. Voir l'article « MacTranslator : a Translator Workstation on a Mac », dans *Language Technology*, n° 10, novembre-décembre 1988, p. 32-34.

4. On regrette que le français n'ait plus guère recours à l'assimilation phonétique chère à nos arrière-grands-parents; on aurait ainsi, et sans effort, à l'instar de *redingote* et de *paquebot*, le *cédérome*.

5. Le numéro spécial de juin-juillet 1989 de la revue *Informateur-Logiciel* fait un tour d'horizon de l'industrie québécoise de l'informatique documentaire.

6. Par saisie, nous entendons ici l'entrée massive et rapide de données, et non la saisie lente par clavier ou par vocoscribe, généralement simultanée à la génération des textes.