Meta
Journal des traducteurs
Translators' Journal

META

La traduction automatique : l'ordinateur au service des traducteurs

Brigitte Roudaud

Volume 37, Number 4, décembre 1992

Études et recherches en traductique / Studies and Researches in Machine Translation

URI: https://id.erudit.org/iderudit/003997ar
DOI: https://doi.org/10.7202/003997ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Roudaud, B. (1992). La traduction automatique : l'ordinateur au service des traducteurs. *Meta*, *37*(4), 828–846. https://doi.org/10.7202/003997ar

Article abstract

Is machine translation a dream or a danger for translators? Such questions are still asked by many translation professionals. In this article we try to show how machine translation is evolving to become a beneficial tool for translators specialised in technical translations. The machine translation systems which are currently on the market are almost always based on computer and linguistic technologies which are now old. Computer science today is highly accessible and flexible; it adapts itself more easily to users. Based on work carried out at GETA (a university laboratory in Grenoble, France), SITE and B'VITAL are currently working on the integration of machine translation in the process of technical document production. We first present different computer and linguistic techniques used; we then talk about evaluation and debugging techniques to be carried out in close collaboration with SITE's professional translators. We conclude with future perspectives represented by the ambitious European project EUROLANG, which has just started.

LA TRADUCTION AUTOMATIQUE : L'ORDINATEUR AU SERVICE DES TRADUCTEURS

BRIGITTE ROUDAUD
B'VITAL (groupe SITE), Grenoble, France

Résumé

La traduction automatique est-elle un rêve inaccessible ou un danger pour les traducteurs ? Voilà des questions que beaucoup de professionnels de la traduction se posent encore. Dans cet article, nous tentons de montrer comment la traduction automatique évolue pour devenir un outil profitable pour les traducteurs spécialisés en traduction technique.

Les systèmes de traduction automatique actuellement commercialisés sont pour la plupart basés sur des technologies informatique et linguistique qui ont vieilli. Aujourd'hui, l'informatique se veut plus accessible, plus souple. Elle s'adapte aux utilisateurs, et non l'inverse. À partir des travaux du GETA (laboratoire universitaire de Grenoble), SITE et B'VITAL travaillent actuellement sur l'intégration de la traduction automatique dans la réalisation des documents techniques. Nous présentons donc tout d'abord les méthodes informatique et linguistique employées, puis les méthodes d'évaluation et de mise au point à mettre en œuvre en étroite collaboration avec les traducteurs professionnels de SITE, enfin nous concluons sur les perspectives d'avenir, concrétisées par le démarrage de l'ambitieux projet européen EUROLANG.

Abstract

Is machine translation a dream or a danger for translators? Such questions are still asked by many translation professionals. In this article we try to show how machine translation is evolving to become a beneficial tool for translators specialised in technical translations.

The machine translation systems which are currently on the market are almost always based on computer and linguistic technologies which are now old. Computer science today is highly accessible and flexible; it adapts itself more easily to users. Based on work carried out at GETA (a university laboratory in Grenoble, France), SITE and B'VITAL are currently working on the integration of machine translation in the process of technical document production. We first present different computer and linguistic techniques used; we then talk about evaluation and debugging techniques to be carried out in close collaboration with SITE's professional translators. We conclude with future perspectives represented by the ambitious European project EUROLANG, which has just started.

INTRODUCTION

Les langues naturelles, véhicules privilégiés de l'information, ont toujours tenu une place importante dans les rapports humains. À l'aube de la «société de l'information» du XXI^{e} siècle, cette place devient capitale pour assurer la compétitivité des entreprises. La documentation commerciale, technique, administrative... est vitale pour assurer la pérennité des produits et des compétences. Une «bonne» documentation — claire, complète, explicite... — est plus stable et plus fiable que l'équipe de développement la plus soudée.

Elle est aussi l'assurance d'une meilleure maîtrise des processus mis en œuvre — ce qui se conçoit bien s'énonce clairement.

Dans la perspective internationale, la traduction de la documentation joue maintenant un rôle particulier. Pour exporter, les entreprises doivent pouvoir fournir leur documentation dans la langue du pays visé — ou, au pire, en anglais — et cela dans les meilleurs délais. Parfois les rédacteurs sont contraints de rédiger dans une langue qui n'est pas la leur, avec des risques dont les conséquences sont plus ou moins importantes au niveau de la compréhension du texte ou de l'appauvrissement de la langue.

Les besoins en traduction deviennent de plus en plus pressants (d'après Gartner Group, le marché mondial de la traduction est d'environ 60 milliards de francs français, c'est-à-dire environ 12 milliards de dollars), dans des domaines parfois très spécialisés demandant des connaissances terminologiques — et même parfois techniques — pointues pour assurer la qualité du texte traduit. Les traducteurs techniques professionnels sont malheureusement en nombre limité, et de nombreuses traductions ne sont jamais effectuées, faute de temps ou de moyens. Pour résorber ce goulot d'étranglement, les professionnels de la documentation en général, et de la traduction en particulier, cherchent des outils qui font encore cruellement défaut. Le traitement automatique des langues (TALN) semble pouvoir apporter des solutions. Les Japonais l'ont bien compris, qui depuis près d'une dizaine d'années font un effort colossal dans ce domaine[1]. Mais les techniques, encore récentes, n'ont pas toujours donné le jour à des outils performants. Ainsi, les premiers systèmes de traduction automatique ont souvent découragé les traducteurs par la mauvaise qualité des traductions.

Faut-il en conclure que le TALN doit encore rester un objet de recherche, sans retombée possible dans l'immédiat — ni même dans un proche avenir ? Ce serait probablement faire erreur. D'ailleurs des outils issus du TALN sont déjà largement utilisés, comme les correcteurs orthographiques, pour ne citer que les plus courants. En ce qui concerne la traduction, la question se pose plutôt en ces termes : les systèmes de traduction automatique — ou d'aide à la traduction — sont-ils actuellement suffisamment «au point» pour être utilisés avec profit par les traducteurs ? Ou le seront-ils dans un proche avenir ?

Dans ce qui suit, nous tentons d'apporter quelques éléments de réponse à ces questions. Pour cela, nous présentons tout d'abord un rapide historique[2] des techniques de TALN en général, et de TA en particulier, en faisant un parallèle avec l'évolution des techniques informatiques. Ensuite, nous exposons notre approche de la traduction automatique, et les solutions que nous préconisons. Puis nous consacrons quelques pages à la question fondamentale du rôle des traducteurs dans la conception et l'évaluation des systèmes de TA. Enfin, nous concluons en exposant les perspectives d'avenir, concrétisées par le démarrage de l'ambitieux projet européen EUROLANG.

1. LA TRADUCTION AUTOMATIQUE ET L'INFORMATIQUE

1.1. LES PREMIERS PAS

Dès l'apparition des premiers ordinateurs, alors que les techniques informatiques étaient encore balbutiantes, les chercheurs ont pensé que de telles machines pourraient traduire aussi bien qu'un traducteur, en beaucoup moins de temps. Il paraissait évident que cela n'était qu'une question de taille de dictionnaires (qu'il faudrait bien sûr énorme) et de puissance de calcul.

Les premières idées étaient issues des travaux informatiques dans le domaine du décryptage des messages codés. De telles méthodes, beaucoup utilisées durant la Seconde Guerre mondiale, semblaient pouvoir être appliquées à la traduction (Weaver

1949/1955), qui s'apparenterait alors à une sorte de décryptage d'une langue vers une autre. Le texte source était considéré comme une sorte de codage, qui par traduction de chacun des symboles (en fait les mots de la langue source) pourrait être directement transposé en langue cible. Cette voie, sans être réellement suivie, a donné naissance aux premiers travaux dans le domaine, privilégiant les méthodes statistiques.

Dès 1954, une équipe de l'université de Georgetown faisait la démonstration d'un système traduisant du russe vers l'anglais. Bien que très limitée, cette démonstration sembla prouver que la traduction automatique était possible.

Les premières expériences étaient basées sur le principe de la traduction mot à mot, où le dictionnaire était l'élément déterminant. De plus, ces systèmes, strictement bilingues, n'étaient fondés sur aucune théorie linguistique.

Les premières améliorations apportées à ces techniques permettaient de modifier localement les positions respectives de certains mots, et de déterminer la traduction du mot par des tests sur son contexte immédiat (les mots voisins). Écrits à l'aide des langages de programmation de l'époque (pas particulièrement adaptés aux problèmes linguistiques), ces systèmes (dont les résultats n'étaient pas si mauvais) péchaient surtout par une maintenance difficile et une évolutivité très limitée. Ces limites semblaient alors difficiles à lever.

1.2. LA LINGUISTIQUE FORMELLE

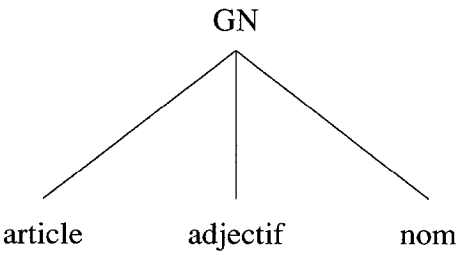
L'évolution de l'informatique, et en particulier l'apparition des premières théories des langages formels (Chomsky 1957) et des premières techniques de compilation, ont alors ouvert de nouveaux horizons. Les langages formels, basés sur la notion de règles, permettaient désormais de décrire de manière simple et déclarative la syntaxe d'un langage de programmation.

Parallèllement, les travaux en linguistique se sont alors concentrés sur la formalisation de la langue naturelle, en vue de réaliser une analyse syntaxique.

Le premier pas franchi fut celui de l'analyse des *syntagmes*, tels que les groupes nominaux (article + adjectif + nom...), les verbes complexes (avec modaux et auxiliaires), etc. Pour décrire un groupe nominal (GN), on peut par exemple écrire une règle telle que :

GN —> article + adjectif + nom

Le groupe nominal ainsi décrit peut se représenter sous la forme de l'arbre :



Ce premier niveau d'analyse, formalisé par les travaux de Chomsky en 1957 est le niveau des «constituants immédiats» définissant les syntagmes.

Un deuxième type de modèle linguistique, formalisé par Gaifman en 1965, définit un niveau d'interprétation supplémentaire, le niveau des relations de dépendance (proche des relations syntaxiques traditionnelles). Présenté en deux mots, ce modèle organise les

constituants de chaque syntagme en relation de dépendance d'un élément principal qui donne sa catégorie au syntagme (l'article dépend du nom, qui lui-même dépend d'un verbe, lorsqu'il en est le sujet, etc.).

Correspondant à la classe des langages hors-contexte, ces deux modèles linguistiques ont donné lieu respectivement aux grammaires syntagmatiques et aux grammaires de dépendance. Des algorithmes de plus en plus performants ont été mis au point pour permettre la production des structures syntagmatiques correspondantes.

Les premiers algorithmes étaient des automates accepteurs, réalisant une analyse ascendante ou descendante, et la structure d'arbre fournie en cas de succès n'était qu'un sous-produit, dont la forme dépendait plus de l'algorithme utilisé que de la description linguistique. Parmi les plus connus, on peut citer les algorithmes de Cocke-Younger-Kasami et d'Earley.

L'adjonction de «traits» véhiculant des informations linguistiques a permis d'augmenter le pouvoir expressif des grammaires hors-contexte (permettant par exemple de vérifier les accords). De nombreux types de grammaires ont ainsi été dérivées des grammaires hors-contexte pures.

Une approche originale de la linguistique formelle a été réalisée par Chomsky au début des années soixante (Chomsky 1957 et Chomsky 1965), ce sont les grammaires transformationnelles. Le point de départ de cette théorie est la démonstration que les grammaires hors-contexte ne pouvaient permettre de représenter et de décrire les phénomènes complexes de la langue. Chomsky propose donc un modèle permettant de dériver la structure de «surface» (c'est-à-dire le texte) de la structure «profonde» (c'est-à-dire son interprétation linguistique) à l'aide de règles transformationnelles. La structure profonde était elle-même décrite par une grammaire hors-contexte. Si cette théorie a pu influencer certains des systèmes existants, aucune application directe n'a pu être réalisée en raison de la trop grande combinatoire inhérente.

Des algorithmes fondés sur un modèle de transduction, donnant la possibilité de contrôler la structure produite par la définition de stratégies et d'heuristiques, ont ensuite vu le jour. Un modèle par transduction prend une structure de données en entrée, et la modifie peu à peu pour produire une structure différente en sortie. On peut mentionner les SYSTÈMES-Q présentés par Colmerauer en 1970, les ATN conçus par Woods en 1968, ou encore le système CETA (*cf.* Vauquois 1967 et 1968), puis le langage ROBRA (*cf.* Boitet *et al.* 1978) mis au point par l'équipe de recherche du GETA.

L'avantage de tels algorithmes est qu'ils permettent de traiter non seulement les aspects syntaxiques, mais aussi, éventuellement, de dériver la structure profonde ou d'effectuer une analyse sémantique.

1.3. LA PROGRAMMATION LOGIQUE ET L'INTELLIGENCE ARTIFICIELLE

Plus récemment (au début des années quatre-vingt) et en partie dérivé des méthodes de programmation logique (Prolog a été conçu au départ pour le TALN), un grand nombre de travaux de recherche s'oriente vers des modèles regroupés sous le terme «grammaires d'unification». De telles grammaires, basées sur l'opération d'unification comme moyen de contrôle du flux d'information, permettent une représentation unifiée des dictionnaires et des grammaires, dans laquelle les informations (lexicales ou grammaticales) sont considérées comme des contraintes de «bonne formation». Parmi les plus connues, citons les grammaires à clauses définies (DCG) de Pereira et Warren (1980), les grammaires lexicales fonctionnelles (LFG) de Kaplan et Bresnan (1981), les grammaires syntagmatiques généralisées (GPSG) de Gazdar et Pullum (1982)...

En parallèle à l'évolution des méthodes de représentation formelle des langues, l'informatique a évolué. L'intelligence artificielle a donné naissance aux systèmes

experts, dont le principe est basé sur le déclenchement de «règles d'inférence» sur une base de connaissances, qui permettent de décrire (ou suivre...) un raisonnement.

L'approche système expert a été tentée par quelques équipes qui ont jugé que la machine ne traduirait correctement qu'en passant par une phase de «compréhension» du texte, par rapport aux connaissances emmagasinées. Deux types de connaissances sont nécessaires : les connaissances linguistiques permettant l'analyse de la syntaxe du texte et les connaissances pragmatiques correspondant au domaine traité par le texte. Parmi les travaux les plus connus, citons le projet KBMT de Carnegie-Mellon qui s'est achevé fin 1989 (*cf.* Nirenburg 1989).

Le point délicat, dans un tel système, est la constitution de la base de connaissance. Les techniques actuelles ne permettent de créer de telles bases, à un coût raisonnable, que pour des «micro-mondes», dans lesquels les connaissances sont bien définies. Malheureusement, d'une part la linguistique ne semble pas faire partie d'un tel micro-monde, d'autre part pour chaque texte traitant d'une nouvelle technique la base de connaissance doit être modifiée.

Un autre axe, plus récent, en intelligence artificielle commence à être exploré, ce sont les réseaux neuronaux. Le principe est la constitution d'un réseau complexe, reliant tous les termes d'une langue, dans lequel les connexions entre les termes le plus souvent mis en rapport dans le langage sont renforcées par rapport aux autres. Un tel réseau se constitue lui-même par apprentissage.

Deux principaux problèmes limitent l'application que l'on pourrait en faire : la taille d'un réseau complet pour une langue et la difficulté (ou l'impossibilité ?) d'effectuer un apprentissage correct.

Les approches intelligence artificielle et grammaire d'unification restent encore actuellement au niveau des centres de recherche. Les systèmes commerciaux actuels, ou en voie de commercialisation, sont basés sur des techniques un peu plus anciennes, mais mieux maîtrisées.

Un autre axe de recherche, exploré par certaines équipes universitaires, est fondé sur des traitements statistiques. Applicables principalement à la résolution des ambiguïtés lexicales, ces méthodes peuvent donner de très bons résultats dans ce domaine. Leur application directe ne s'est pas encore concrétisée dans les systèmes de TAO industriels.

1.4. LE CONTEXTE INDUSTRIEL

Pendant que la recherche évolue au niveau de la linguistique informatique, les industriels quant à eux ont réalisé les avantages que pouvait leur apporter une terminologie spécialisée bien gérée. On a pu ainsi voir de grands industriels, faute de produits commerciaux adaptés, réaliser leur propre système maison (parfois ensuite commercialisé) pour gérer leur terminologie. Citons les systèmes TEAM de SIEMENS, PHENIX de SITE...

Les instances européennes et internationales, sensibles au problème de l'évolution des langues et en particulier de l'évolution des langues de spécialités, cherchent à encourager la standardisation terminologique, au niveau des termes, des méthodes et des outils. De grandes banques terminologiques ont ainsi été créées, dont les premières dès le début des années soixante. Parmi les banques actuelles, citons EURODICAUTOM de la CEE, NORMATERM de l'AFNOR, la Banque de terminologie du Québec, TERMIUM, etc.

L'évolution récente de l'informatique ouvre de nouveaux horizons pour la gestion lexicale en général. Ainsi, l'industrialisation des bases de données relationnelles d'une robustesse suffisante pour accueillir la somme des informations linguistiques indispensables aux applications en TALN permet d'envisager la création de «bases lexicales». Depuis peu, de nombreux projets, visant à définir un modèle conceptuel cohérent, ont vu

le jour. Ces études ont pour but essentiel la définition d'un modèle générique pouvant accueillir les données lexicales de toutes les applications.

Les premières bases lexicales, reposant sur ces principes de généricité et d'évolutivité, commencent à apparaître. Le groupe SITE, pour sa part, a réalisé pour le CNET (FRANCE TÉLÉCOM) et commercialise *Le Lexicaliste*, base lexicale générique et configurable, dans un environnement graphique moderne et ergonomique.

Enfin, sur le plan du matériel, l'arrivée sur le marché des machines RISC proposées par l'ensemble des constructeurs permet, à coût égal, d'augmenter considérablement la puissance disponible. Les prévisions pour 1994 laissent penser que l'utilisateur pourra alors bénéficier d'une configuration fournissant de 50 à 100 MIPS (million d'instructions par seconde) pour un prix n'excédant pas 25 000 FF (5 000 $ US). Dans ces conditions, en extrapolant les chiffres actuels, une page de 250 mots pourrait être traduite par la machine en 25 secondes, et son coût informatique (en tenant compte de l'amortissement de la machine sur trois ans et d'une production de 1 000 pages par an) n'excèderait pas 10 FF (2 $ US).

Pour conclure sur l'évolution des techniques, la traduction automatique (TA), bénéficiant maintenant des nouvelles techniques informatiques et d'une puissance des machines accrue, peut enfin entrer dans l'âge adulte. Depuis quelques années déjà, des systèmes commerciaux ont prouvé que la TA pouvait, dans certaines conditions, apporter une aide aux industriels. Les notions fondamentales d'ergonomie et de facilité d'emploi des machines ouvrent peu à peu la porte du monde informatique et de la traduction automatique à Monsieur Tout-le-Monde.

2. UNE APPROCHE INDUSTRIELLE DE LA TAO

Dans une optique industrielle, le but principal d'un système de TAO est de permettre aux traducteurs de gagner du temps et de l'argent, et donc de pouvoir traduire un plus grand volume de documents. Pour cela, un certain nombre de facteurs apparaissent prépondérants :

- la qualité des traductions,
- la facilité d'emploi du système,
- la robustesse du système,
- l'efficacité du traitement automatique.

Par ailleurs, au niveau des développements, il est nécessaire de minimiser les coûts de développement et de maintenance.

2.1. QUALITÉ DES TRADUCTIONS

Étant donné l'état actuel des techniques, tous les spécialistes du domaine reconnaissent qu'il est impossible d'obtenir une qualité parfaite, sauf peut-être dans un domaine et sur une syntaxe très limités (comme le prouve l'exemple du système TAUM-MÉTÉO, qui traduit parfaitement 90 % des bulletins météorologiques au Canada, les autres 10 % étant rejetés par le système et traduits «à la main»...).

L'approche interactive, où le système pose des questions (plus ou moins pertinentes) au traducteur ou au rédacteur pendant le processus de traduction automatique, peut permettre d'améliorer la qualité des traductions. Malgré tout, les systèmes basés sur ce principe ne donnent pas un résultat parfait.

Certains systèmes ont privilégié les techniques de pré-édition. La syntaxe et le vocabulaire utilisés sont alors contrôlés sur le texte source. Les phrases sortant du cadre sont rejetées. Généralement, il est alors demandé à l'auteur de les modifier. On peut donc

tout de suite identifier un premier problème : l'auteur (ou au moins une personne habilitée à faire de telles modifications) doit intervenir, il faut donc pouvoir le joindre... De plus, pour que la qualité des traductions soit excellente, la syntaxe et le vocabulaire doivent être très stricts, ce qui est difficile à faire accepter aux rédacteurs. Lorsque la syntaxe est moins stricte, les traductions ne peuvent pas être garanties «parfaites».

Une solution intermédiaire, entre la syntaxe contrôlée ou très réduite et l'acceptation de n'importe quel type de textes, consiste à avoir un système orienté pour la traduction d'un type de textes. Basée sur une notion étendue des sous-langages (*cf.* Kittredge et Lehrberger, 1982), cette approche considère que les textes peuvent être classés suivant deux critères :

- le vocabulaire utilisé,
- la typologie des textes (c'est-à-dire le style des textes, au sens large, regroupant la syntaxe, l'usage des tournures stylistiques, l'usage de la ponctuation, l'organisation textuelle, etc.).

Chacun reconnaît, par exemple, qu'il est vain de chercher à traduire automatiquement de la littérature, alors que les textes très techniques posent généralement moins de problèmes (il y a moins de sous-entendus, moins de figures de styles...).

La limitation du vocabulaire intervient au niveau du vocabulaire «général» (des mots courants tels que *maman* ne se rencontrent pas dans les textes techniques et tous les «niveaux de langage» ne sont pas admis) comme de la terminologie correspondant au domaine traité.

D'un point de vue industriel, la part la plus importante du marché de la traduction est occupée par la traduction technico-commerciale (qui représente 40 % des traductions effectuées) et les chiffres prouvent que la demande dans ce domaine ne cesse de croître (de 10 à 15 % par an). C'est pourquoi, fort de sa propre expérience, SITE prévoit d'orienter ses efforts dans le sens de la traduction des documents techniques.

Ce type de choix a deux avantages principaux :

- la réduction des coûts de développement et de maintenance, au niveau des grammaires comme au niveau des dictionnaires ;
- la possibilité d'améliorer (par des choix pertinents) la qualité des textes correspondant à la typologie et au domaine ciblés.

D'après ce que l'on a vu précédemment, pour obtenir un document technique commercialisable, une phase de révision (postédition) est indispensable, quel que soit le système utilisé. Notons que les services de traduction procèdent tous à une étape de révision des traductions, par un autre traducteur, avant de fournir le résultat au client. Pour l'anecdote, certains traducteurs trouvent ce type de révision délicat... il n'est pas question de mettre en cause le style du traducteur initial. L'ordinateur lui n'est pas sensible aux critiques.

Pour que le système soit tout de même efficace, il est donc nécessaire de s'assurer que la révision d'un texte est sensiblement plus courte qu'une traduction complète du même texte. Des essais effectués à SITE sur les résultats de traduction automatique fournis par le système ARIANE (logiciel développé par le GETA de Grenoble) sur le couple français-anglais dans le domaine de l'aéronautique (développé par B'VITAL, filiale de SITE) ont montré que les temps de révision étaient environ 50 % plus courts que les temps de traduction.

Plus que des critères de qualité linguistique, difficiles à déterminer, ces chiffres obtenus sur des textes «réels» (dont SITE était chargé de la traduction) ont permis de confirmer le groupe SITE dans sa volonté d'investir dans le domaine.

2.2. FACILITÉ D'EMPLOI

Pour que le système soit facile à utiliser, il est essentiel :

- qu'il soit un outil agréable (ergonomie),
- qu'il soit intégré à l'environnement habituel de production des traductions (intégration).

L'ergonomie implique d'une part l'accès simple et rapide à toutes les fonctionnalités dont le traducteur a besoin, d'autre part un environnement permettant un apprentissage rapide (donc un environnement «naturel», faisant référence à des notions usuelles pour les traducteurs et fournissant des fonctions d'aide).

Les techniques actuelles de réalisation d'interfaces homme-machine graphiques permettent de proposer un tel environnement.

L'intégration dans l'environnement du traducteur est plus délicate à réaliser, dans la mesure où de nombreux environnements existent. Au niveau des textes eux-mêmes se pose le problème du format, bien connu de tous ceux qui ont un jour essayé de récupérer un document... car il est généralement écrit avec un traitement de texte différent de celui (bien meilleur) que l'on a l'habitude d'utiliser.

Des traitements automatiques sont possibles, pour passer d'un format à un autre, même s'ils nécessitent parfois des développements informatiques très spécifiques.

2.3. ROBUSTESSE DU SYSTÈME DE TA

Dans la pratique, les textes techniques «réels» posent toujours des problèmes que les systèmes n'ont pas prévu (mots nouveaux, utilisation peu classique de la ponctuation ou des éléments entre parenthèses, phénomènes linguistiques non traités...).

Ceci est d'autant plus vrai que l'on privilégie l'aspect «qualité-prix». En effet, on peut choisir de ne pas traiter certains phénomènes linguistiques rares lorsqu'ils nécessitent une mise en œuvre coûteuse (recherche d'une solution linguistique pertinente et développement des grammaires correspondantes). Il en est de même au niveau du vocabulaire.

De nombreux systèmes de laboratoire ne produisent aucune solution en cas de problème (si l'analyse échoue, par exemple).

Il existe une autre raison d'échec, plus insidieuse. En effet, certains systèmes (y compris des systèmes commerciaux actuels) échouent lorsque les phrases sont trop longues (les limites sont généralement de l'ordre de 20 à 30 mots). Comme on le verra par la suite, ce phénomène est lié à la combinatoire de la langue.

Un deuxième critère de robustesse est la qualité de la traduction en cas de problème. En effet, lorsque l'analyse échoue sur une phrase donnée, généralement une partie de la phrase a tout de même pu être correctement analysée. On aimerait pouvoir utiliser ce résultat partiel pour la suite du traitement, garantissant ainsi qu'une partie de la phrase, au moins, sera correctement traduite.

La robustesse du système se situe donc à deux niveaux :

- produire toujours au moins une solution, quel que soit le texte source,
- prévoir des méthodes de rattrapage, permettant de garantir un minimum de qualité, même en cas d'échec d'une partie des traitements linguistiques.

2.4. EFFICACITÉ DU TRAITEMENT AUTOMATIQUE

On entend ici par efficacité la possibilité de fournir suffisamment de traductions aux réviseurs... En effet, aussi bonnes que soient les traductions, si elles demandent des temps de traitement très longs, les traducteurs préféreront traduire directement eux-mêmes.

En pratique, on peut résoudre en partie ce problème d'une part en jouant sur la puissance et le nombre des machines de traduction, et d'autre part en effectuant les traductions automatiques la nuit et le week-end. Mais des temps d'attente trop longs sont malgré tout un problème, en particulier lorsqu'il s'agit du coût et de l'amortissement des machines.

Ce problème n'est pas aussi simple à résoudre qu'il pourrait y paraître. En effet, la plupart des systèmes (et en particulier des systèmes commerciaux actuels) sont basés sur des systèmes de règles non déterministes. Dans un tel système, toutes les analyses possibles doivent être effectuées, les unes après les autres, le choix de la meilleure étant éventuellement fait a posteriori, en fonction de critères linguistiques spécifiques.

Le problème est que les langues sont intrinsèquement ambiguës... et que de nombreuses options sont ainsi généralement essayées. Lorsque les traitements qui en découlent deviennent tellement énormes qu'ils «effondrent» les performances des machines, on parle de l'explosion combinatoire du modèle.

En pratique, l'explosion est directement liée au nombre de mots dans la phrase. Pour simplifier, plus il y a de mots et plus il y a de combinaisons possibles de ces mots.

Certains systèmes, pour éviter des temps d'attente trop longs dans ce genre de cas, préfèrent provoquer un arrêt brutal des traitements au bout d'un certain laps de temps. L'utilisateur a alors l'impression que le système impose une limite sur la taille des phrases qu'il accepte...

Comme on vient de le voir, l'informatique (la puissance des machines et l'efficacité des algorithmes) n'est pas seule en cause dans les problèmes de performance ; le modèle linguistique utilisé joue lui aussi un grand rôle.

2.5. MINIMISATION DES COÛTS DE DÉVELOPPEMENT ET DE MAINTENANCE

Parmi les différents moyens utilisés pour minimiser les coûts de développement et de maintenance, le premier est la définition et l'utilisation de langages (ou formalismes) spécialisés, grâce auxquels les traitements linguistiques sont décrits, et à partir desquels des techniques algorithmiques sophistiquées peuvent être mises au point. Cela paraît plus ou moins évident pour la plupart des développeurs actuels, en raison de l'évolution de l'informatique et de l'apparition de langages de programmation «évolués», mais l'était beaucoup moins au début des travaux en la matière.

De plus, comme en informatique «classique», une méthodologie (de «génie linguistique») est un atout supplémentaire pour aider aux développements et à la maintenance. Dans cet optique, deux «couches» de grammaires peuvent être définies :

- les grammaires descriptives, permettant aux linguistes de spécifier aussi naturellement que possible les phénomènes linguistiques traités et les structures linguistiques qui leur sont associées ;
- les grammaires applicatives, qui décrivent les traitements effectués pour atteindre l'objectif fixé par les spécifications.

Au niveau de la maintenance, le système doit permettre le repérage facile des erreurs, leur correction ainsi que l'ajout éventuel de nouveaux traitements. Il est fréquent, dans les systèmes de TALN, de provoquer des «phénomènes de bord» lors de la maintenance. Cela se traduit généralement par le fait que des phrases jusqu'alors correctement

traitées ne le sont plus. Le GETA, qui avait tout d'abord mis au point un système basé sur un système de règles de type déclaratif, s'est heurté à ce type de problème et le système ARIANE actuel a été mis au point en tenant compte de l'expérience acquise à ce niveau.

Par ailleurs, pour un système de TAO, la gestion du multilinguisme est un facteur important pour réduire les coûts de développements. En effet, les premiers systèmes de TA développés étaient purement bilingues. Dès que l'on changeait l'une des langues (source ou cible), tout était à refaire.

Conscients de ce problème, les chercheurs ont alors conçu l'idée d'un système basé sur le principe d'une analyse et d'une génération **monolingues**. Classiquement deux approches ont été suivies :

- système avec langage pivot,
- système par transfert.

Dans le premier cas, l'analyse doit conduire à une représentation du «sens du texte», totalement indépendante de la langue source. Parfois certains systèmes utilisent une langue naturelle comme langage pivot, c'est le cas par exemple du projet DLT au Pays-Bas (*cf.* Witkam 1984) qui utilise ainsi l'espéranto (en considérant que c'est effectivement une langue naturelle).

Dans le deuxième cas, l'analyse, plus ou moins profonde suivant les systèmes, est ensuite suivie d'un transfert, qui transforme la structure source, obtenue en analyse, en une structure compatible avec la langue cible.

La première méthode est fort attrayante a priori dans un contexte multilingue. Le GETA, qui avait tout d'abord tenté cette approche, a mis rapidement en évidence des problèmes importants :

- la difficulté de définir un langage pivot universel non ambigu ;
- le problème, non résolu par le langage pivot, de l'analyse qui doit toujours être parfaite pour atteindre le langage pivot ;
- l'impossibilité d'utiliser des informations contrastives, permettant de simplifier les traitements ou de guider les choix de syntaxe en génération.

Enfin, étant donné la quantité et la qualité nécessaires au niveau des informations lexicales dans un système de TAO, la constitution et la maintenance des dictionnaires posent un problème de coût critique. Pour réduire au minimum ces coûts, la solution base lexicale générique (présentée précédemment) est généralement considérée comme la plus performante.

De plus, la possibilité de récupérer dans une telle base des données terminologiques existantes (par exemple à partir des bases terminologiques «maison») est capitale pour la constitution de lexiques fiables. Par ailleurs, l'ergonomie des postes de saisie, modification et consultation de la base lexicale doit être particulièrement soignée (et adaptée).

2.6. LES SOLUTIONS ADOPTÉES

Globalement, les solutions que nous préconisons, pour le traitement des langues naturelles en général et pour la TAO en particulier, sont donc, au niveau informatique comme au niveau linguistique :

- utilisation de langages spécialisés pour le développement des applications linguistiques suffisamment puissants pour permettre le traitement du plus grand nombre de langues possible, et en particulier des langues les plus couramment utilisées dans l'industrie ;

- utilisation d'une base lexicale multilingue indépendante des applications linguistiques ;
- traduction «par transfert», les phases d'analyse et de génération étant totalement monolingues ;
- analyse syntaxo-sémantique profonde, limitant ainsi autant que possible la phase de transfert ;
- utilisation d'un transducteur (c'est-à-dire un outil de manipulation et de transformation des structures linguistiques — généralement représentées sous forme d'arbre —), permettant de limiter la combinatoire du modèle linguistique grâce à l'utilisation de méthodes heuristiques, et assurant la production d'au moins une traduction ;
- spécifications linguistiques dans un formalisme le plus neutre possible par rapport aux applications et aux théories ;
- structure linguistique «multiniveau», permettant de «limiter les dégâts» en cas d'échec partiel de l'analyse ;
- approche pragmatique, basée sur le rapport qualité/prix et sur la définition d'une typologie des textes techniques ;
- importance de l'ergonomie et de la qualité informatique du système ;
- intervention des utilisateurs (c'est-à-dire les traducteurs) dans les phases de spécification et de test.

3. UNE FRUCTUEUSE COLLABORATION AVEC LES TRADUCTEURS

Les traducteurs ont longtemps considéré la TAO avec un regard ambivalent. D'une part, ils critiquaient la mauvaise qualité des traductions brutes produites par l'ordinateur, et, d'autre part, ils craignaient une trop grande perfection qui risquerait d'amoindrir leur rôle, et même de les mettre au chômage.

De plus, bien que l'ordinateur ait maintenant envahi le monde industriel, nombre de traducteurs travaillent encore avec le dictaphone et le papier (fiches terminologiques, dictionnaires, versions papier de la traduction après saisie par une secrétaire...). Ces méthodes, fort louables, évoluent peu à peu, et les nouvelles générations, familiarisées avec l'informatique dès leur plus jeune âge, trouvent plus facile de travailler leurs traductions directement avec un traitement de texte et accès direct à une base terminologique intégrée. Chez SITE, les traducteurs travaillent déjà tous dans un environnement largement informatisé.

Au niveau de la qualité, les progrès de la linguistique formelle permettent d'atteindre dorénavant un niveau suffisant pour ne pas rebuter complètement les traducteurs. Par expérience, les traducteurs ont aussi appris à limiter leurs espérances, l'ordinateur traduit moins bien qu'eux et c'est très bien ainsi. Restait à prouver qu'un système de traduction automatique pouvait leur apporter quelque chose dans le travail quotidien.

Pour situer le problème, il est bon de rappeler très succinctement quelques aspects de la documentation technique. Le style des documents ainsi que le vocabulaire général employé sont généralement pauvres. Certaines phrases reviennent de façon répétitive dans les textes. Après traduction, les documents évoluent souvent (surtout les documents de maintenance industrielle de matériels) et les «révisions» doivent à nouveau être traduites. La technique décrite peut être très pointue et utiliser une terminologie peu familière au traducteur (pour résoudre ce dernier point, les services de traduction «spécialisent» souvent leurs traducteurs). Dans un domaine donné, la terminologie (source et/ou cible) peut varier avec le client (suivant le constructeur, on trouve *train d'atterrissage* ou *atterrisseur*). Après traduction, les textes sont systématiquement relus par un

autre traducteur (une ou deux fois, au moins, suivant le client) qui propose éventuellement des corrections.

Tout cela rend souvent les traductions techniques difficiles et fastidieuses. L'utilisation de systèmes gérant les révisions et la terminologie améliore les conditions de travail. Dans ce cadre, couplé avec des fonctionnalités adaptées, le système de TAO peut devenir un outil apprécié. Déchargé d'une partie des tâches fastidieuses, le traducteur peut se concentrer sur le style et la compréhensibilité du texte.

Pour arriver à un tel objectif, la collaboration des traducteurs est indispensable pour :

- fixer la limite inférieure de qualité linguistique acceptable ;
- déterminer les éventuelles erreurs de traduction que le traducteur «ne supporte pas» et chercher autant que possible à les éliminer ;
- prévoir les fonctionnalités techniquement réalisables et réellement utiles au traducteur ;
- améliorer l'ergonomie du poste du traducteur, en se référant à des notions qui lui sont familières (et qui ne sont peut-être pas familières aux linguistes et aux informaticiens).

Au niveau de la qualité linguistique, seule une phase de test et d'évaluation par les traducteurs peut apporter les réponses à ces questions. Une première évaluation, limitée, a déjà été réalisée sur le système français-anglais existant (développé par B'VITAL), révélant que la qualité actuelle des traductions est acceptable. Des tests plus poussés sont encore nécessaires pour déterminer les points à améliorer en priorité.

Au niveau du poste du traducteur, des développements sont encore nécessaires pour améliorer l'intégration du système dans le processus documentaire ainsi que les fonctionnalités et l'ergonomie du poste. Les projets de SITE vont permettre de résoudre ce problème crucial. En effet, dans le projet EUROLANG, la qualité de l'environnement de travail sera l'un des soucis constants des développements informatiques. La participation de traducteurs professionnels est l'une des garanties de succès du projet, quant à son adéquation par rapport aux besoins des utilisateurs.

4. EUROLANG: LE PROCHE AVENIR

Le projet EUROLANG, qui a débuté en novembre 1991, est un projet EUREKA ambitieux, d'un budget d'environ 500 millions FF (100 millions $ US) sur trois ans. Il a pour objectif essentiel le développement en trois ans d'une boîte à outils TALN, dont la première application sera un système de traduction assistée par ordinateur (TAO) basé sur l'état de l'art en informatique comme en linguistique. Le produit de TAO sera commercialisé à l'issue du projet.

4.1. UNE STRATÉGIE DE L'INNOVATION

Le groupe SITE (SONOVISION ITEP Technologies), filiale du holding CORA-REVILLON, est chef de file européen en ingénierie de la documentation technique. Par son métier, SITE est donc directement engagé dans tous les aspects du traitement de la langue : rédaction, traduction, terminologie, gestion de nomenclature, hypertexte, etc.

Parmi les différentes activités liées à la gestion de la documentation technique, la traduction prend une part de plus en plus importante, dans un monde où les échanges internationaux vont croissant. SITE gère actuellement l'un des plus grands centres de

traduction européens, avec une production annuelle de plus de 100 000 pages pour des clients tels que IBM, DEC, NUM, AÉROSPATIALE, DASSAULT AVIATION...

Impliqué dans la TAO depuis 1982, SITE s'est investi plus nettement depuis le rachat de sa filiale spécialisée B'VITAL. Dans le cadre d'une convention avec le ministère français de l'Industrie, SITE a tout d'abord testé et industrialisé la technologie ARIANE sur le couple français-anglais. Cette technologie, conçue et développée par le GETA, laboratoire universitaire de Grenoble, est reconnue par les experts mondiaux du domaine. Elle a d'ailleurs inspiré les travaux japonais pour la réalisation de produits basés sur les techniques de seconde génération.

Pour assurer l'avenir industriel de cette technologie, le groupe SITE estime que des actions sont nécessaires, principalement :

- l'écriture d'un nouveau noyau, visant une meilleure portabilité et une meilleure rentabilité, et incluant des innovations technologiques souhaitables ;
- une interconnexion avec des dictionnaires multilingues existants;
- une amélioration du formalisme linguistique, pour minimiser le coût d'adjonction d'un nouveau couple de langues ;
- une distribution mondiale du produit par des sociétés spécialisées.

La Commission des Communautés Européennes, dans le cadre du projet EUROTRA, a permis à l'Europe de faire progresser le savoir linguistique de ses meilleures équipes universitaires. La Commission est favorable à l'organisation d'une bonne synergie entre ses propres actions à moyen et à long terme et les projets industriels à court terme pour les raisons suivantes :

- les produits japonais de seconde génération risquent de prendre une place prédominante en Europe à l'horizon de 1995, faute d'une réponse industrielle rapide ;
- une synergie analogue coordonnée par le MITI a fait la preuve de sa pertinence pour le couple de langues japonais-anglais ;
- la maîtrise des échanges multilingues est un objectif stratégique pour l'Europe.

Dans cette optique, sans utiliser les développements réalisés dans le cadre du projet EUROTRA, le projet EUROLANG bénéficiera de l'expertise et du savoir-faire d'une partie des équipes de développement d'EUROTRA, attirée par la perspective de développer un système opérationnel.

Siemens Nixdorf, qui développe et commercialise le système METAL, est particulièrement intéressé par la définition d'une plateforme européenne commune en matière de TALN et prend une part active au projet.

Dans cette perspective, le produit EUROLANG pourrait s'inscrire comme une solution d'avenir pour le TALN européen, regroupant des technologies industrielles éprouvées et bénéficiant d'une expérience commerciale unique.

4.2. LES PARTENAIRES

Pour atteindre l'objectif fixé dans le projet EUROLANG, des industriels et des universitaires européens ont donc décidé de mettre en commun leur ressources techniques, humaines et commerciales.

Cinq pays participent à cet effort : la France, l'Allemagne, l'Angleterre, l'Espagne et l'Italie. Dans chaque pays, des équipes universitaires, dont la plupart ont participé au projet EUROTRA, et des industriels spécialisés dans le domaine ou motivés par des besoins réels ont rejoint le consortium.

En France, outre SITE (maître d'œuvre), le consortium regroupe Cap Gemini Innovation, le GETA, le LADL de M. Gross, le CNET (FRANCE TÉLÉCOM) et l'utilisateur potentiel Matra Space Marconi.

En Allemagne, SIEMENS NIXDORF, dont les compétences en matière de TAO sont mondialement reconnues et qui souhaite étendre ses applications en matière de TALN, participe de façon active au projet, en apportant en particulier une compétence rare dans le domaine de la commercialisation d'un système de TA. L'IAI (institut de recherche ayant participé à EUROTRA) et Krupp Industries (utilisateur potentiel) font également partie du projet.

En Angleterre, Rank Xerox, qui a développé sur SYSTRAN une interface homme-machine de qualité, interviendra au niveau des interfaces, et les deux équipes spécialisées en linguistique informatique, de l'université d'Essex et de l'UMIST (Manchester) mettront à profit dans le projet EUROLANG l'expérience acquise au sein du projet EUROTRA.

En Espagne, BDE, spécialiste de documentation, et deux équipes universitaires, basées à Barcelone et dont l'une a participé à EUROTRA, se sont joints au projet.

En Italie enfin, plusieurs compagnies commerciales et équipes de recherche, dont certaines ont participé à EUROTRA, ont rejoint le consortium. Citons Lexicon, Thamus, Gruppo Dima et les universités de Pise (Professeur Zampolli) et de Salerne.

4.3. LES OBJECTIFS INDUSTRIELS

L'objectif premier du projet est la réalisation d'un système de TAO commercial, dont la version 1.0 offrira 10 couples de langues en 1995 : 8 couples entre l'anglais et les quatre autres langues du projet (français, allemand, espagnol et italien), plus les 2 couples français<—>allemand. Ces 10 couples de langues sont en effet jugés stratégiques pour les besoins de la communication technique européenne.

Tout au long du projet, les principaux choix techniques seront guidés par les motivations industrielles, à savoir la prédominance des besoins des utilisateurs. La présence de tels utilisateurs dans le consortium (SITE, maître d'œuvre, étant l'un d'eux) est un gage de succès quant à l'adéquation du produit à ce niveau.

En pratique, les maîtres-mots en matière de produits industriels sont : ergonomie, robustesse, efficacité, adéquation et souplesse d'emploi du produit, garantie sur la maintenance et l'évolutivité, portabilité... Dans cette optique, si la qualité linguistique des traductions est un facteur prépondérant, elle n'est pas l'unique contrainte.

Dans le domaine du TALN, un autre facteur important entre en jeu : le coût de constitution des dictionnaires nécessaires. La réutilisabilité des données lexicales ainsi que des mécanismes permettant de minimiser les coûts de constitution des dictionnaires sont donc deux paramètres dont il faut tenir compte pour un produit industriel.

Pour répondre à ces besoins, les partenaires ont défini une politique commune basée sur la recherche des meilleures solutions pratiques aux vrais problèmes rencontrés par les utilisateurs. Cette démarche nécessite en particulier la mise en œuvre en priorité des techniques haut de gamme éprouvées dans le domaine.

Enfin, la mise en place, parallèlement au projet, d'un club utilisateur, garantira l'adéquation du système aux différents besoins du marché. Le club utilisateur, constitué d'un nombre volontairement limité d'industriels européens, aura un droit de regard sur l'évolution du projet et un rôle de consultant pour la définition des besoins. De plus, les premières versions des différents produits seront en priorité mises à la disposition des membres du club.

Parallèlement à ce premier objectif à court terme, et dans une optique industrielle réaliste, les partenaires du projet visent à préparer l'avenir à moyen terme en matière de TALN. Pour cela, deux types d'actions sont prévues :

- dans sa version 1.0, le produit EUROLANG possèdera toutes les qualités nécessaires à sa future mise à niveau par rapport à l'évolution des techniques de TALN ;
- une action de type recherche et développement sera maintenue dans le projet, visant à étudier de nouvelles solutions pour les produits futurs. En particulier, la mise en œuvre de nouvelles techniques basées sur l'unification sera un des axes de recherche privilégié.

La présence de groupes de recherche au sein du projet est donc un facteur indispensable pour assurer la pérennité de l'offre EUROLANG.

4.4. LES CHOIX TECHNIQUES

Les choix techniques sont basés sur l'état de l'art en matière de logiciels et sur les contraintes imposées par les choix industriels présentés ci-dessus. Très généralement, des notions telles que portabilité, extensibilité, modularité, réutilisabilité... et utilisation de standard (UNIX, C/C++, SGBD Relationnels/SQL, X11/MOTIF ou WINDOWS, SGML...) sont des critères de performance industrielle auxquels les partenaires du projet sont attachés. Tous les développements seront donc réalisés en tenant compte de ces critères.

L'**extensibilité** et la **portabilité** du produit seront assurées au niveau logiciel par le développement en C/C++ d'une **boîte à outils**, à partir de laquelle les différentes applications TALN pourront être réalisées. L'optique boîte à outils présente de nombreux avantages par rapport aux objectifs industriels du projet. En effet, une telle boîte à outils peut facilement évoluer par l'ajout de nouveaux éléments. Ceci assure que l'évolution des techniques linguistiques pourra être facilement mise à profit à moyen ou long terme. Par ailleurs, un mécanisme souple permettant d'interconnecter les différents outils disponibles donne des garanties quant à la **modularité** et à l'**ouverture** des applications développées à l'aide de la boîte à outils, ainsi qu'à la **réutilisabilité** des différentes briques de bases ainsi constituées (un «lemmatiseur», par exemple, est indispensable dans de nombreuses applications linguistiques).

L'**ergonomie** des langages et des interfaces homme-machine sera un des soucis premiers des concepteurs. Elle permet d'assurer à l'utilisateur le confort nécessaire à une utilisation harmonieuse des produits mis à sa disposition. Dans l'optique de la boîte à outils, l'ergonomie, garantissant une meilleure productivité, se situe à plusieurs niveaux :

- au niveau des développeurs d'une application linguistique, telle qu'un système de TAO ;
- au niveau des utilisateurs des applications linguistiques développées, comme les traducteurs utilisant l'application TAO.

Dans cet esprit, toutes les techniques modernes seront mises en œuvre pour la réalisation des outils (syntaxe des langages de programmation simple et homogène, environnement de développement fournissant des outils de mise au point évolués...) et des interfaces homme-machine (multi-fenêtrage, souris, icônes, boutons, affichage graphique, fonction d'aide en ligne...).

L'exploitation d'une **base lexicale multilingue et générique**, pour la gestion des données lexicales (dictionnaires, bases terminologiques), permettra la récupération de

dictionnaires existants et la réutilisabilité des données lexicales pour différentes applications linguistiques. Un poste d'enrichissement lexicographique souple et adapté est de plus indispensable pour améliorer la productivité des lexicographes.

Enfin, pour favoriser la réutilisabilité des outils et même des développements linguistiques, il est indispensable de prévoir dès le départ la gestion du **multilinguisme**. Dans le projet EUROLANG, le multilinguisme se situe à plusieurs niveaux :

- l'expressivité des outils (c'est-à-dire leur capacité de donner les moyens aux linguistes de décrire comme ils le souhaitent des phénomènes linguistiques quelconques), qui est une garantie de pouvoir traiter une nouvelle langue ;
- la gestion des données lexicales dans une base lexicale multilingue ;
- la gestion des textes, en particulier au niveau du problème des différents jeux de caractères ;
- au niveau de l'application TAO, l'approche «par transfert», qui permet d'écrire des analyses et des générations totalement monolingues et réutilisables.

4.5. LA BOÎTE À OUTILS EUROLANG

Le principe de la boîte à outils repose sur la définition d'un certain nombre d'outils et d'une API (*application programming interface*), pour gérer la communication entre les outils. L'environnement de développement fourni avec la boîte à outils offrira des mécanismes souples pour interconnecter les différents outils en vue de la réalisation d'une application donnée.

Parmi les outils, des priorités de développement ont été définies, de manière à produire d'abord les briques de base indispensables à la première application visée, à savoir le système de TAO. Les premiers éléments seront basés sur des techniques d'ores et déjà validées et offrant toutes les garanties de succès. Citons par exemple :

Une base lexicale multilingue qui sera l'outil générique, souple et puissant, grâce auquel seront gérées toutes les données lexicales multilingues. Cette base sera réalisée à partir de l'expérience acquise par les différents partenaires dans le domaine. Elle prendra en compte, autant que faire se peut, les recommandations émises par les différents projets européens en la matière (GENELEX, MULTILEX, EUROTRA 7...).

Un transducteur d'arbre décoré. Cet outil sera basé sur le langage ROBRA, utilisé dans le système ARIANE, et étendu par de nouvelles fonctionnalités, si nécessaire.

Un langage de traitement morphologique permettant d'analyser ou de générer les mots (au cours de la phase de spécification, on précisera s'il faut des outils différents pour l'analyse et la génération). Ce langage permettra de décrire comment les mots sont formés, à partir, par exemple, de préfixes, bases, suffixes dérivationnels et flexionnels. Pour permettre la gestion du multilinguisme (c'est-à-dire pouvoir réaliser un analyseur pour «n'importe quelle langue»), il est nécessaire de fournir un formalisme simple permettant d'écrire des règles spécifiant le découpage des mots. Un tel mécanisme doit, par exemple, pouvoir manipuler les mots composés allemands en les découpant correctement.

Un langage de traitement des expressions : l'analyse des expressions (ou idiomes) est généralement délicate dans la plupart des langues, d'autant plus qu'il existe de nombreuses expressions non figées (variables et/ou non connexes). Même les expressions les plus simples, comme par exemple *pomme de terre* varient (au pluriel *pommes de terre*). Certains cas sont toutefois plus complexes, comme par exemple *mettre le moteur en marche* où l'expression à reconnaître est *mettre en marche*, ou encore *j'ai pris mes jambes à mon cou* où l'expression à reconnaître est *prendre ses jambes à son cou*.

Un langage d'analyse à contexte limité d'emploi très simple. Un tel langage pourrait par exemple être utilisé pour résoudre les ambiguïtés lexicales, avant le processus

d'analyse, de manière à limiter la combinatoire en analyse. Ce langage pourrait aussi être utilisé pour traiter des phénomènes syntaxiques locaux bien définis, comme par exemple les dates, ou même pour réaliser une première étape de construction syntaxique simple.

4.6. LE PRODUIT TAO EUROLANG

Au niveau du développement du produit de TAO EUROLANG à partir de la boîte à outils, deux axes se distinguent: les développements informatiques et les développements linguistiques.

4.6.1. DÉVELOPPEMENTS INFORMATIQUES

Ces développements concernent principalement les aspects interface homme-machine au sens large. En effet, plusieurs types d'interfaces (et les fonctionnalités associées) sont indispensables:

l'atelier TAO qui permet aux linguistes de développer les grammaires et les dictionnaires dans un environnement ergonomique adapté à leurs besoins. Un tel atelier prendra en charge un certain nombre de problèmes pratiques du type enchaînement des outils spécifiques à l'application de TAO, gestion des grammaires et des dictionnaires, gestion des versions de programmes et des données textuelles, etc.;
l'environnement de gestion des dictionnaires permettant d'accéder à la base lexicale par une interface spécialisée pour l'enrichissement lexical et d'extraire de la base les informations nécessaires à l'application de TAO;
le poste du traducteur / réviseur, poste de l'utilisateur final de l'application, fournissant un environnement agréable, en fonction des besoins et des souhaits exprimés par les utilisateurs. Des fonctions spécifiques pour la pré-édition et la postédition sont prévues (accès à la base, aide au marquage des hors-texte, aide à l'accentuation des textes source, visualisation synchronisée des textes source et traduit, édition de statistiques de traitement, etc.). Pour une meilleure intégration de ce poste dans la chaîne de production de la documentation, des traitements automatiques permettront de transformer les textes sources de leur format initial vers le format pivot SGML qui sera défini pour les textes, et inversement pour les textes traduits;
le poste d'analyse de corpus utile principalement pour permettre la recherche automatique d'expressions, dans un corpus donné, et pour gérer les phrases répétitives, évitant ainsi de traduire plusieurs fois une même phrase. Cette dernière fonctionnalité est appréciable non seulement pour la traduction d'un document donné (dans la langue technique certaines phrases reviennent très souvent), mais aussi pour la traduction des versions successives d'un même document, dont une partie du texte n'a pas été retouchée. La recherche automatique des expressions, basée sur les travaux récents de différents partenaires, est une aide précieuse pour la constitution d'une terminologie spécialisée.

4.6.2. DÉVELOPPEMENTS LINGUISTIQUES

Classiquement, les développements linguistiques comprennent deux aspects: les dictionnaires (généraux et terminologiques) et les grammaires.

Les dictionnaires seront constitués à l'aide de la base lexicale, autant que possible par récupération et enrichissement des données existant chez les différents partenaires. Il est prévu d'obtenir des dictionnaires généraux d'environ 50 000 termes pour chaque langue (et couple de langues).

Les grammaires couvriront une large part des phénomènes linguistiques des différentes langues traitées. Dans un souci de performance et de qualité, elles seront orientées

vers le traitement de la documentation technico-commerciale (manuels d'entretien, de maintenance, d'utilisation, etc.).

En ce qui concerne les développements linguistiques monolingues, chaque pays est responsable des développements concernant sa langue. Pour les développements bilingues, chaque pays est responsable des transferts vers sa langue, sauf pour le cas de la langue anglaise pour laquelle une partie des transferts est confiée à d'autres pays, de manière à équilibrer les charges.

Aussi, pour assurer la convergence des développements linguistiques, une législation linguistique, définie en début de projet, règlera entre autres le délicat problème de la cohérence des structures linguistiques interfaces (par exemple, un résultat d'analyse doit fournir les informations nécessaires au transfert, qui lui-même doit fournir une structure linguistique interprétable par la génération). La législation linguistique doit aussi fixer les limites des couvertures grammaticale et lexicale du système de TAO.

CONCLUSION

L'informatique est une science ou une technique jeune, encore pleine de promesses. Les langues, toujours en évolution, ne sont ni science ni technique. L'art d'écrire ou de parler est souvent mal maîtrisé par les hommes eux-mêmes, et les documents techniques en sont souvent une criante illustration. Les techniques de traitement automatique des langues cherchent à concilier la logique informatique et la créativité linguistique. Le but de la machine devient peu à peu ce qu'il aurait toujours dû être, une aide précieuse à la créativité.

Comme les autres applications possibles de ces techniques, la traduction automatique, libérée de la contrainte inaccessible de la perfection, peut devenir un outil efficace avec l'aide des traducteurs professionnels. Les chiffres sont là, qui prouvent qu'avec un minimum de qualité les traductions fournies par la machine peuvent faciliter la tâche des traducteurs et leur faire gagner du temps, même si le coût prohibitif d'un système d'une qualité suffisante avait jusqu'à présent retardé le développement de cette technique.

Conscients des enjeux économiques et culturels associés aux industries de la langue, et encouragés par les progrès récents, les industriels et les instances gouvernementales tentent de trouver des solutions économiquement viables. Le groupe SITE, dont le permier métier est la documentation technique sous toutes ses formes, prend une part active à cette démarche industrielle. Le projet EUROLANG en est la concrétisation la plus parlante.

Gageons que les efforts actuels donneront des résultats à court et moyen terme. À plus long terme, la recherche continue, améliorant les techniques, les méthodes, testant des idées nouvelles... Les industries de la langue ont encore leur avenir devant elles.

Notes

1. Citons en particulier le projet EDR (1986-1995) d'un budget de 1,5 milliard de francs (300 millions $ US), et le projet ATR, d'un budget de 4 milliards de francs (800 millions $ US) sur 15 ans.
2. L'historique présenté est non seulement rapide mais très simplifié. Toutes les théories, méthodes, systèmes... ne peuvent être présentés, ni même cités, en si peu de pages. Nous nous en excusons auprès des experts.

BIBLIOGRAPHIE

BACHUT, D. *et al.* (1991) : «Industrialisation d'un système de TAO français-anglais pour la documentation technique», *Génie Linguistique, 91*, Versailles.

BOITET, C. *et al.* (1982) : «ARIANE-78 : An Integrated Environment for Automated Translation and Human Revision», *COLING-82*, Prague, 1982.

BOITET, C. (1986): «The French National MT-Project: Technical Organization and Translation Results of CALLIOPE-AERO», *IBM Conference on Translation Mechanization*, Copenhague.
BOITET, C. (1986): «Current Machine Translation Systems Developed with GETA's Methodology and Software Tools», *ASLIB*, London, 1986.
BOITET, C. *et al.* (1978): «Manipulation d'arborescence et parallélisme: le système ROBRA», *COLING-78*, Bergen.
BOITET, C. (1987): «Current State and Future Outlook of the Research at GETA», *MT Summit*, Hakone.
CHANDIOUX, J. (1976): «MÉTÉO: un système opérationnel pour la traduction des bulletins météorologiques destinés au grand public», *Meta*, 21-2.
CHAPPUY, S. (1983): «Formalisation de la description des niveaux d'interprétation des langues naturelles», Thèse de 3e cycle informatique, Grenoble.
COLMERAUER, A. (1970): «Les systèmes-Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur», Université de Montréal.
GAIFMAN, H. (1965): «Dependency Systems and Phrase Structure Systems», *Information and Control, 8*.
GAZDAR, G. et G. PULLUM (1982): «GPSG: A Theoretical Synopsis», Indiana University Linguistics Club.
GAZDAR, G. *et al.* (1985): «Generalized Phrase Structure Grammar», Oxford, Basil Blackwell.
GROSS, M. et A. LENTIN (1967): «Notions sur les grammaires formelles», Paris, Gauthier-Villard.
HUTCHINS, W.J. (1986): «Machine Translation: past, present, future», Chichester, Ellis Horwood Series in Computer and their Applications.
KAPLAN, R. et J. BRESNAN (1982): «Lexical-functional grammar: A formal system for grammatical representation», Bresnan J. (Ed.), *The Mental Representation of Grammatical Relations*, Cambridge, MIT Press.
KITTREDGE, R. et J. LEHRBERGER (1982): «Sublanguages: study of language in restricted semantic domains», Berlin, de Gruyter.
NIRENBURG, S. (1989): «KBMT-89 — A Knowledge-based MT Project at Carnegie Mellon University», MT Summit II, Munich.
PEREIRA, F. et D. WARREN (1980): «Definite Clause Grammars for Natural Language Analysis — A Survey of the Formalism and a Comparison with Augmented Transition Networks», *Artificial Intelligence, 13*.
SABAH, G. (1988): «L'intelligence artificielle et le langage», Paris, Hermes.
VAUQUOIS, B. (1967): «Le système de traduction automatique du CETA», Congrès d'EREVAN, URSS.
VAUQUOIS, B., «Structures profondes et traduction automatique — Le système du CETA», *Revue roumaine de linguistique, 13*.
VAUQUOIS, B. et S. CHAPPUY (1985): «Static Grammars: a Formalism for the Description of Linguistic Models», *International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, Colgate University.
WEAVER, W. (1955): «Translation», *Machine Translation of Language*, Locke et Booth, Technology press of MIT and Wiley and Sons, New York, (première parution en 1949).
WITKAM, A.P.M. (1984): «Distributed language translation, another MT system», *International Seminar on Machine Translation*, Cranfield, England.
WOODS, W. (19 «Transition Network Grammars for Natural Language Analysis», *CACM*, 13/10.