

LA SYNERGIE ENTRE THAM, RÉSEAU ET TA COMME FACTEUR DE PROGRÈS THÉORIQUES ET PRATIQUES EN TAO

The synergy between MAHT, the Net and MT as a factor for progress in automated translation

Christian BOITET

GETA, CLIPS, IMAG (UJF & CNRS),
150 rue de la Chimie, BP 53
38041 Grenoble Cedex 9, France
Christian.Boitet@imag.fr

Résumé

Les systèmes de TA et d'aide au traducteur (THAM) sont de plus en plus utilisés conjointement en traduction professionnelle lourde, mais leur intégration laisse encore beaucoup à désirer. La recherche d'un meilleur couplage mène à de nouvelles architectures lexicales et grammaticales. Grâce au réseau, on pourra aussi faire bénéficier les traducteurs occasionnels d'outils professionnels de THAM. En conjuguant le couplage TA/THAM et l'utilisation intensive du réseau, on pourra enfin construire des systèmes de TA fondée sur le dialogue (TAFD) de grande taille, utilisables depuis des micros de bas de gamme pour rédiger dans sa langue et être traduit dans une ou plusieurs autres, avec la possibilité de transmettre exactement le sens désiré dans un format "auto-explicatif".

Mots-clés: TA, THAM, TAFD, TA/THAM, télé-TAO, base lexicale intégrée, serveur lexical.

Abstract

MT systems and computer tools for human translators (MAHT) are increasingly used jointly in heavy professional translation, but their integration is still wanting. The quest for a better coupling leads to new lexical and grammatical architectures. Through the net, it will also be possible to make professional MAHT tools usable by occasional translators. Conjugating a better MT/MAHT coupling and the intensive use of the network will also make it possible to build large Dialogue-Based MT (DBMT) systems, usable from low-end micros to write in one's language and be translated into one or several others, with the possibility to transmit exactly the desired meaning in a "self-explaining" format.

Keywords: MT, MAHT, DBMT, MT/MAHT, tele-MAT, integrated lexical data base, lexical server.

Introduction

Les outils de THAM utilisés en traduction professionnelle lourde utilisent un lemmatiseur et une mémoire terminologique pour suggérer des équivalents, et une "mémoire de traduction" ("mémoire phraséologique"), composée de fragments bilingues alignés (phrases, titres, énoncés), qui permet de retrouver un fragment déjà traduit, identique ou très similaire au fragment à traduire.

Pour certains types de textes, comme des versions successives de fichiers d'aide en ligne, ou des avertissements successifs en cas de crise (problèmes routiers, grèves, catastrophes naturelles...), on arrive à des taux de répétition très importants. Par contre, d'autres, qui semblent plus simples, comme les bulletins météorologiques, sont certes construits sur un vocabulaire restreint, et utilisent un répertoire limité de constructions syntaxiques, mais ne présentent pas une bonne répétition au niveau des phrases complètes¹. L'utilisation d'un système de TA est alors une meilleure solution.

Or, dans la plupart des flots de traduction, on trouve à la fois des parties adaptées aux mémoires de traduction, des parties plus adaptées à la TA, et des parties provisoirement ou définitivement non traitables automatiquement, comme des listes de nouveaux produits, des phrases d'un nouveau sous-langage, ou des slogans publicitaires. Longtemps séparés, les outils de TA et d'aide au traducteur (THAM) sont donc de plus en plus utilisés conjointement. Ainsi, TMTM d'IBM est couplé avec LMT, et EuroLang OptimizerTM avec LogosTM. Ce couplage consiste

¹ D'après des mesures récentes de J. Chandioux, on trouve environ 1.000 nouvelles structures de surface (suites de catégories) par jour dans les 60.000 mots (200 bulletins) traduits quotidiennement d'anglais en français par le système de TA METEO.

simplement à soumettre à un serveur de TA les énoncés ou phrases trop distants du candidat le plus proche trouvé dans la mémoire de traduction, ou tous. On pourrait aussi n'envoyer un fragment au serveur de TA que si tous ses mots ou presque sont connus du système de TA. Mais, en pratique, on ne le sait pas, car les deux systèmes sont indépendants.

Ce manque d'intégration n'est pas seulement mauvais du point de vue économique (on envoie à la TA des fragments qui vont échouer et donc produire de mauvaises suggestions), mais pénalisant pour les traducteurs, qui peuvent faire évoluer les dictionnaires de la THAM, mais pas ceux de la TA, dont les résultats n'auront pas la cohérence terminologique souhaitable.

Un autre problème de ces outils de THAM lourde est que leur nature et leur prix les réservent à un usage professionnel : d'après EuroLang, il faut traduire environ 800 à 1.000 pages par an à la fois pour amortir un tel outil et le rendre efficace (par croissance des mémoires), ce qui le rend inutilisable par des traducteurs occasionnels isolés.

Une meilleure synergie entre THAM et TA, au niveau non seulement de leur usage, mais de leur construction, peut résoudre ces deux problèmes, et permettre le partage de tels systèmes par des traducteurs occasionnels, grâce au réseau, sans qu'ils aient à acquérir une licence, et de façon que chaque communauté constitue ses mémoires sur le serveur.

Dans les deux premières parties, nous montrons comment ces idées peuvent faire progresser la méthodologie de la construction des systèmes de TAFD² classiques", en construisant une base lexicale d'où tous les dictionnaires de THAM et de TA seront extraits, et en adaptant un système comme EuroLang Optimizer à l'usage occasionnel via InterNet.

Reste le problème de la traduction lourde vers un grand nombre de langues, et des communications multilingues entre individus. Le paradigme de la TAFD (traduction automatique fondée sur le dialogue) apporte une réponse théorique, mais la construction et la maintenance de systèmes de TAFD "pour tous" posent de gros problèmes de coût et de simple faisabilité : les dictionnaires devront être très grands, et où trouver les équivalents, si les utilisateurs sont monolingues ?

Dans la troisième partie, nous montrons comment, en conjuguant le couplage TA-

/THAM et l'utilisation intensive du réseau, on pourra construire des systèmes de TAFD de grande taille, utilisables depuis des micros de bas de gamme. Les progrès envisageables consistent à transformer la base lexicale en un serveur incrémental accédé par réseau, et la base phraséologique "plate" en une base structurée dans un format "auto-explicatif", ce qui permettra, d'affiner la recherche d'énoncés voisins, et, sans connaître les langues cibles, de sélectionner la proposition de traduction correspondant au sens désiré.

1. Synergie THAM/TA

1.1 Motivations

Coupler un système de THAM et un système de TA développés et évoluant indépendamment pose des problèmes de cohérence lexicale, de qualité et de faisabilité. Pour le premier, nous rappelons plus loin une solution connue, mais pas encore appliquée..., car elle suppose résolus les deux autres.

En ce qui concerne la qualité, un constructeur de systèmes de TA ne peut suivre correctement qu'un nombre limité de systèmes à la fois. On fait donc évoluer la plupart des systèmes de TA en jouant uniquement sur le contenu et les priorités des dictionnaires utilisateurs, le plus souvent uniquement terminologiques, ce qui limite fortement la qualité maximale accessible.

Pour l'augmenter, il faudrait pouvoir faire évoluer le cœur du système, c'est à dire adapter les dictionnaires généraux et les heuristiques syntaxiques et sémantiques. Mais les constructeurs de systèmes de TA répugnent à transférer à leurs clients tout ou partie de leur savoir faire. Dans l'état actuel des choses, la seule solution est de demander au constructeur de développer une version de base de son système de TA adaptée à la typologie visée, puis de la faire évoluer normalement.

Apparaît alors le problème de faisabilité. Comment en effet financer un tel effort, qui se chiffre à 30—50 hommes*mois pour un système de qualité, sachant que ledit système ne sera peut-être amorti que sur la fraction du flot de traduction inadaptée à la mémoire de traduction et adaptée à la TA, par exemple 30 à 40% dans des pages Web ?

1.2 Construire un système de TA tout en traduisant par THAM

La solution que nous proposons ici est purement méthodologique.

Acquérir un système existant au meilleur prix et le faire évoluer n'est pas satisfaisant,

² TA fondée sur la langue, à dictionnaires et règles.

pour les raisons exposées plus haut. Chercher une application susceptible d'être traitée en totalité par la TA, puis construire un système de TA adapté, en 2-3 ans, sans rien traduire, est en général difficile, risqué, et peu réaliste. En effet, il faut à la fois que la tâche de traduction puisse attendre que le système de TA soit opérationnel, qu'on puisse financer son développement, et qu'on soit sûr que son efficacité sera suffisante pour gagner assez sur les traductions pour l'amortir dans un délai raisonnable, sachant que, dans la plupart des cas, on sera amené à mettre en place un système de THAM pour la fraction du flot de traduction qui se révélera inadaptée à la TA.

Nous proposons alors d'inverser les étapes, c'est à dire d'installer un système de THAM, de commencer tout de suite à traduire, et de développer en parallèle le système de TA, en faisant participer les traducteurs à la construction des dictionnaires, et en finançant la construction du système de TA par le gain obtenu après sa mise en service, en couplage avec le système de THAM.

Prenons l'exemple d'un volume de 50.000 à 60.000 pages standard de 250 mots. Cela représente environ 66.700 à 80.000 heures de travail sans THAM. Avec THAM, en supposant 20% de coïncidences exactes et 40% de coïncidences approchées, on tombe à 33.000 ou 40.000 heures.

Supposons par exemple une équipe de 4 développeurs informaticiens et linguistes, et de 6 traducteurs. En utilisant le système de THAM, les traducteurs construisent la mémoire terminologique. Ils passent de plus environ 30% de leur temps à indexer les propriétés syntaxiques et sémantiques de ces termes dans les dictionnaires de TA.

Supposons qu'il faille 12 mois pour construire la première version opérationnelle du système de TA, à partir d'une version de base existante. Les traducteurs auront alors passé 7.000 heures à traduire ou réviser, et produit 10.000 pages. Supposons qu'on remplace alors deux développeurs par deux traducteurs, et que les huit traducteurs continuent à indexer les nouveaux termes dans les dictionnaires de TA, à raison de 30% de leur temps. Ils travailleront au total de 13.000 à 16.300 heures, soit de 13,5 à 16,5 mois.

Les deux tableaux suivants illustrent cette idée sur plusieurs cas de figure. On voit cependant que, avec les chiffres retenus dans l'exemple précédent, il faudrait pouvoir compter sur un flot de traduction d'au moins 60.000 pages pour que la combinaison THAM/TA devienne intéressante. Elle le deviendrait pour un volume plus faible si la fraction adaptée à la TA croissait, couvrant éventuellement tout ou partie de celle adaptée à la mémoire de traduction.

	h/p ou %	Hyp. 1	Hyp. 2	Hyp. 3	Hyp. 4	Hyp. 5
Pages		10.000	15.000	40.000	50.000	60.000
Traduction brute humaine	1 h	10.000	15.000	40.000	50.000	60.000
Révision	0,33 h	3.333	5.000	13.333	16.667	20.000
TH (heures)	1,33 h	13.333	20.000	53.333	66.667	80.000
Correspondances exactes	20%	2.000	3.000	8.000	10.000	12.000
Correspondances approchées	40%	4.000	6.000	16.000	20.000	24.000
Autres	40%	4.000	6.000	16.000	20.000	24.000
Traduction brute humaine	1 h	4.000	6.000	16.000	20.000	24.000
Révision	0,33 h	2.667	4.000	10.667	13.333	16.000
THAM (heures)	0,67 h	6.667	10.000	26.667	33.333	40.000
Traduction brute humaine	1 h	0	0	0	0	0
Révision	0,33 h	2.667	4.000	10.667	13.333	16.000
THAM/TA (heures)	0,27 h	2.667	4.000	10.667	13.333	16.000

Mois écoulés	142,33 h	12	18	24	30	36
Pages produites/traducteur (10 traducteurs THAM)	214	2.562	3.843	5.124	6.405	7.686
Pages produites en THAM	149	1.793	0	0	0	0
6 traducteurs à 70%	897	10.760	0	0	0	0
Pages produites en THAM/TA	374	0	2.242	4.484	6.725	8.967
8 traducteurs à 70%	2.989	0	17.934	35.868	53.802	71.736
Pages produites (THAM/TA)		10.760	28.694	46.628	64.562	82.496

Dans ce scénario, l'utilisateur participe à la construction du système de TA, en affectant deux personnes à l'équipe de développement, par exemple un informaticien et un ingénieur linguiste. Il les garde ensuite pour maintenir et améliorer le système.

Dans l'affaire, le constructeur du système de TA de base n'est pas perdant, bien au contraire. En effet, il évite d'avoir à supporter une obligation de résultat sur de multiples systèmes, tout en pouvant fournir de la formation, des outils annexes, et la maintenance de la partie purement informatique de l'ensemble (environnements de développement et d'utilisation, langages spécialisés).

1.3 Architecture lexicale intégrée pour la THAM et la TA

Si l'on veut que les traducteurs indexent, il faut sans doute limiter la finesse de description linguistique, ou en tout cas organiser la description des propriétés lexicales à plusieurs niveaux. C'est d'ailleurs ce qui est fait par Sharp pour le système DUET : le système de base ne contient qu'assez peu de traits sémantiques, mais on peut rajouter à volonté dans les dictionnaires des codes liés au domaine, que les grammaires utilisent indirectement, donc sans devoir être modifiées.

Pour résoudre le problème de la cohérence, il faut que les informations lexicales soient centralisées dans une base de données lexicales multilingues (BDLM) unique, d'où on extrait les dictionnaires du système de THAM et du système de TA. Cette idée de base lexicale multi-application est une des motivations de la plupart des projets récents sur les dictionnaires.

Cependant, elle n'a à notre connaissance été mise en œuvre que partiellement. On a en effet construit des bases lexicales qui sont spécifiques de la TA, tout en étant indépendantes d'un système particulier³, mais pas de bases lexicales intégrant les termes et les informations nécessaires à la TA (termes généraux aussi bien que terminologie, catégories morpho-syntaxiques, cadres prédicatifs, rections syntaxo-sémantiques, traits sémantiques, dérivations, codes d'acceptions...) et celles utiles à la THAM (très peu d'informations grammaticales, peu ou pas de vocabulaire général, mais des définitions et des exemples).

³ Par exemple, la base BDTAO de B'VITAL/SITE du projet Ariane/aéro/F-E, ou la base lexicale du système MU/Majestic du KICST à Tokyo.

La construction de telles bases suppose une organisation informatique très ouverte (Boitet & al. 1986a, Sérasset 1994a), délicate mais possible à mettre en place avec les techniques actuelles (Sérasset 1994d). Si elles n'existent pas encore, c'est à notre avis qu'on ne peut les construire que dans un contexte où les techniques de THAM et de TA sont conjuguées étroitement, comme dans le scénario décrit ci-dessus.

2. THAM haut de gamme pour traductions non professionnelles grâce au réseau

2.1 Motivations

Dans beaucoup de pays non anglophones, on cherche à promouvoir l'usage de la langue nationale dans les domaines scientifiques et techniques, tout en reconnaissant l'utilité de l'anglais. Mais la traduction et l'interprétation professionnelles sont hors de prix pour les acteurs concernés, et ne sont pas ou très peu subventionnées.

Le cas de la francophonie, en France au moins, est typique à cet égard. La communauté française de la recherche est loin de refuser de communiquer en français. Mais les chercheurs, pris par la multiplicité de leurs tâches : recherche, enseignement, administration, jurys, colloques et congrès, ne produisent trop souvent pas de versions françaises de leurs articles, communications et rapports⁴.

C'est tout à fait compréhensible, car on peut estimer qu'un chercheur sachant bien l'anglais passe entre 1h et 2h par page standard de 250 mots⁵ pour traduire ce qu'il a écrit⁶, soit entre 10h et 20h pour un petit article de colloques (10 pages standard), en thème comme en version.

Exhorter les auteurs à traduire eux-mêmes leurs textes ne mène donc à rien. Si on ne peut financer la traduction professionnelle, on pourrait au moins financer des outils d'aide à la traduction personnelle comme de l'équipement de base, tout comme on finance les matériels et logiciels de bureautique.

⁴ Quant à l'interprétation, on l'exige souvent des organisateurs de colloques, sans leur accorder une subvention augmentée d'autant !

⁵ Une page d'un article comme celui-ci correspond à un peu plus de 2 pages standard.

⁶ Un professionnel passe 1h à traduire et 20mn à réviser, pour produire de bonnes traductions.

Mais les technologies efficaces d'aide à la traduction reposent sur les répétitions de termes et de phrases dans de gros volumes de traduction. Elles sont donc actuellement réservées aux traducteurs professionnels.

2.2 Le projet "Montaigne"

Cependant, plusieurs produits commerciaux actuels sont organisés de façon que les traducteurs, travaillant en groupe, partagent un serveur de "prétraduction" sur un réseau local, et utilisent sur leur poste de travail des aides interactives intégrées au traitement de texte. C'est le cas d'EuroLang Optimizer, autour duquel nous avons conçu et proposé le projet "Montaigne" (en attente).

EuroLang Optimizer fournit une aide à la traduction à partir de six langues sources (français, anglais, allemand, espagnol, italien, néerlandais) vers onze langues cibles (les six langues sources ainsi que le danois, le finnois, le norvégien, le portugais, et le suédois). À l'usage, on nous dit que la vitesse de traduction croît au moins d'un facteur 1,6, et peut croître d'un facteur 3,3 s'il y a beaucoup de répétitions⁷.

L'architecture est de type client-serveur. Un texte à traduire est prétraité ("prétraduit") sur le serveur, qui construit un "kit" associé, constitué du dictionnaire local des termes du document, des phrases de la base égales ou similaires aux phrases du document, et éventuellement de traductions brutes automatiques. Le traducteur reçoit le tout sur son poste client. Grâce à une interface très bien intégrée au logiciel de traitement de texte (Word.6™ sur PC/Windows ou FrameMaker™ sur station Unix), il construit sa traduction à partir des suggestions venant du kit. Ce faisant, il enrichit le dictionnaire local, actif immédiatement. Les ajouts terminologiques et les traductions venant des traducteurs sont ensuite intégrés aux dictionnaires et mémoires de traduction du serveur.

L'efficacité de l'aide fournie dépend cruciallement de la taille des dictionnaires et des mémoires de traductions. Un chercheur isolé ne peut donc utiliser ce système avec profit.

L'idée de base du projet est de mettre un tel serveur, gratuitement, à la disposition de la communauté scientifique. En retour, les auteurs, traduisant leurs documents de façon beaucoup plus efficace, accepteraient que leurs enrichissements terminologiques et leurs

traductions finales (de phrases) soient intégrés aux bases du serveur, et réutilisés pour la recherche et/ou dans des applications.

L'adaptation de cet outil de THAM à des communautés nombreuses mais dispersées, où chaque individu ne produit que de faibles volumes de traduction, consiste ici à :

- 1) mettre en place un serveur de "prétraduction" accessible via tous les réseaux par simple courrier électronique.
- 2) modifier le logiciel "client" (EO-client) fournissant les aides interactives à la traduction sur l'ordinateur de l'auteur-traducteur pour (1) qu'il n'y ait plus besoin de licence spécifique, et (2) qu'il renvoie automatiquement au serveur le kit modifié une fois la traduction terminée.

Voici un scénario un peu plus détaillé montrant la faisabilité de la chose. Si le chercheur n'a pas le logiciel client installé sur son PC, il envoie au serveur un message particulier⁸, et reçoit un formulaire d'authentification et de renseignements divers (institut, domaine de recherche, autorisation de communiquer sa terminologie et sa phraséologie, etc.). EO-client lui est alors envoyé par le réseau ou par la poste sur CD-ROM, avec sa documentation en ligne.

À partir de son poste de travail, et par simple courrier électronique, le chercheur envoie le texte qu'il veut traduire au serveur distant. Pour cela, il utilise EO-client, qui lui demande de fixer les paramètres de la prétraduction (domaines, langues,...), et envoie un message auquel il attache un fichier codé dans un format adéquat.

Le serveur effectue la "prétraduction" du texte (fabrication du kit à partir des dictionnaires et des mémoires de traduction), puis renvoie au chercheur le kit et un jeton permettant d'utiliser EO-client pour une durée limitée et seulement sur ce document.

Le chercheur traduit alors en bénéficiant des aides interactives de EO-client, totalement intégré à Word™. Ce faisant, le kit s'enrichit des phrases traduites et des ajouts au dictionnaire local faits par le chercheur.

Quand la traduction est finie, EO-client renvoie le kit au serveur. Les administrateurs du réseau valident ensuite les termes introduits et les phrases traduites, qui viennent ainsi enrichir la base de données du serveur.

⁷ 1,6 correspond à un gain de productivité de 30%, et 3,3 à 70%, observé chez certains utilisateurs.

⁸ du genre de ceux demandés par tous les "listserv".

Les communautés de recherche utilisant Montaigne-EO pourraient ainsi bénéficier d'une aide efficace à la traduction, reposant sur des dictionnaires et des mémoires de traduction multi-domaines de taille importante, et suivant naturellement la création terminologique et phraséologique.

La contribution "en retour" des utilisateurs est primordiale pour la création des bases. En effet, la création des dictionnaires par des professionnels serait d'un coût rédhibitoire⁹, et ne pourrait jamais être à jour dans tous les domaines. La création a posteriori de mémoires phraséologiques par alignement serait aussi beaucoup plus difficile et beaucoup plus onéreuse.

Elle permettra aussi beaucoup d'apports intéressants, tant vers la recherche que vers l'industrie, sous forme de terminologie, de phraséologie et de corpus, via des serveurs nationaux et européens, en respectant les divers droits éventuels des auteurs et autres intervenants.

On peut envisager de mettre à la disposition des chercheurs :

- les bases terminologiques (termes alignés) ;
- les bases phraséologiques (phrases alignées) ;
- les bases textuelles (textes alignés).

En ce qui concerne la légalité de ces mises à disposition :

- Pour les bases terminologiques, il ne devrait pas y avoir de problème, car il s'agit de contributions sans caractère original¹⁰.
- De même, il ne devrait pas y avoir d'obstacle majeur à la communication des bases phraséologiques à fins de recherches, puisqu'il s'agit de phrases isolées.
- Par contre, la situation est plus délicate en ce qui concerne les textes eux-mêmes, car le droit moral de l'auteur s'exerce. Dans le processus de traduction présenté plus haut, les textes ne font normalement que transiter par le serveur, et n'y sont pas stockés. Il faudrait donc un développement spécifique pour pouvoir en stocker certains, avec l'accord explicite des auteurs. Les textes (sources et traduits) ainsi mis à disposition

par leurs auteurs seraient alors transmis aux serveurs de corpus, à fins d'études¹¹.

De telles bases intéressent les chercheurs tant pour des études "amont" que pour la construction de systèmes concrets. L'évolution de la terminologie et de la phraséologie pourrait ainsi être étudiée par les chercheurs en linguistique, terminologie, et TALN.

Les bases terminologiques pourraient aussi servir de point de départ à la constitution de dictionnaires de TAFD, dans les domaines considérés. Nous détaillons cela plus bas.

Les bases terminologiques et phraséologiques pourraient aussi servir à construire des systèmes d'indexation automatique.

Les bases textuelles pourraient enfin servir de matériau de base pour des recherches sur la catégorisation des textes, l'alignement, l'extraction de terminologie bilingue, et la traduction automatique "par analogie".

En ce qui concerne les retours vers l'industrie, il peut y avoir des conditions plus strictes liées à l'usage commercial, mais il ne faudrait pas que le coût soit trop élevé. En ingénierie linguistique, la francophonie accuse en effet un retard considérable par rapport à d'autres communautés linguistiques. Il est important que, comme au Japon (projet EDR), les ressources lexicales puissent être mises à disposition du plus grand nombre possible d'acteurs économiques.

3. Synergie THAM/TA et réseau pour la TAFD

3.1 La TAFD : intérêt et problème de construction

De plus en plus, nous désirons rédiger dans notre langue, et transmettre nos textes à l'étranger, qu'il s'agisse de messages électroniques, de lettres, d'articles, de manuels techniques, voire de livres. Contrairement à ce que d'aucuns prédisaient il y a une cinquantaine d'années, l'internationalisation croissante ne s'est pas accompagnée d'une uniformisation linguistique vers l'anglais, mais au contraire d'un renforcement considérable de l'usage scientifique et technique de langues déjà importantes de ce point de vue, comme le japonais, et d'une promotion volontariste de

⁹ On cite le chiffre de 20mn par terme bilingue.

¹⁰ Dans le très rare cas d'un néologisme que l'auteur voudrait protéger, il lui suffirait de ne pas le mettre dans le lexique du kit et d'utiliser la fonction Chercher/Remplacer de Word !

¹¹ Au niveau français, il s'agit des serveurs du PRC-CHM et de l'AUFELF•UREF/CNRS. Au niveau européen, il s'agit de l'ELRA, dont la vocation est justement de distribuer les ressources linguistiques en prenant en compte tous les aspects légaux.

bien d'autres, comme le malais-indonésien ou l'arabe, pour les amener au même niveau.

À notre sens, et même si l'anglais continue de se renforcer en tant que volapük scientifique moderne, cette évolution se poursuivra, les langues étant, pour reprendre une expression de C. Hagège dans un article paru dans *Le Monde* début 1990, les "drapeaux des identités nationales".

Il ne s'agit pas seulement de politique, mais d'efficacité. Dans les projets coopératifs européens (Esprit, Eureka), par exemple, la communication est gênée par la nécessité de lire et d'écrire en anglais. Pour la grande majorité des participants, lire en anglais pose des problèmes de compréhension et prend trop de temps. Quant à écrire, si même c'est envisageable, le résultat est souvent difficile à comprendre, voire illisible. De grandes sociétés comme Thomson chiffrent à plusieurs dizaines de MF par an les pertes dues aux problèmes de communication multilingue.

Les trois types de TAO "classique" ne peuvent évidemment répondre à ce nouveau besoin. En effet, la TAO du veilleur, sans préédition ni postédition, ne peut donner une qualité suffisante, et la TAO du réviseur comme la TAO du traducteur s'adressent par définition à des spécialistes au moins bilingues, et non à des rédacteurs supposés ne connaître aucune des langues cibles, ou au plus une, et ce imparfaitement.

L'idée de la TAFD date des années soixante (Kay 1973), et a été incorporée à plusieurs maquettes ou prototypes dans les années soixante-dix à quatre-vingt (Brown 1989, Chandler & al. 1987, Melby & al. 1980, Sadler 1989, Tomita 1986, Whitelock & al. 1986, Wood 1989). Si ces travaux n'ont pas donné lieu à des systèmes utilisables en pratique, c'est à notre avis que les dialogues devaient être conduits par des spécialistes¹², que la couverture linguistique était trop limitée, et que l'on ne disposait pas encore d'environnements interactifs conviviaux.

La méthodologie s'est affinée ces dernières années. Tout d'abord, l'utilisateur envisagé n'est plus un spécialiste, mais un rédacteur, ou plutôt un *auteur* (Boitet & al. 1994a, Boitet & al. 1994b, Brown & al. 1990, Huang 1990, Maruyama & al. 1990, Somers & al. 1990, Wehrli 1991, Wehrli 1992). Nous

préférons ce dernier terme. D'un côté, en effet, "auteur" est moins restrictif que "rédacteur" : un auteur est quelqu'un qui veut créer un texte, et peut le faire en l'écrivant, en le dictant, ou encore en le construisant interactivement. De l'autre, "auteur" est plus restrictif que "rédacteur", "locuteur", ou "commentateur". "auteur" désigne en effet quelqu'un qui désire créer un produit final "propre", alors que les autres termes peuvent renvoyer à des personnes désirant seulement produire un message écrit ou parlé de façon spontanée, en vue d'une communication immédiate, et non disposées à mener un dialogue éventuellement lourd pour rendre leur message "propre".

D'autre part, l'informatique personnelle a fait des progrès gigantesques. On dispose maintenant d'ordinateurs personnels très puissants et bon marché, d'environnements conviviaux, de l'intégration du multimédia, et d'outils de télécommunication permettant le recours à des serveurs.

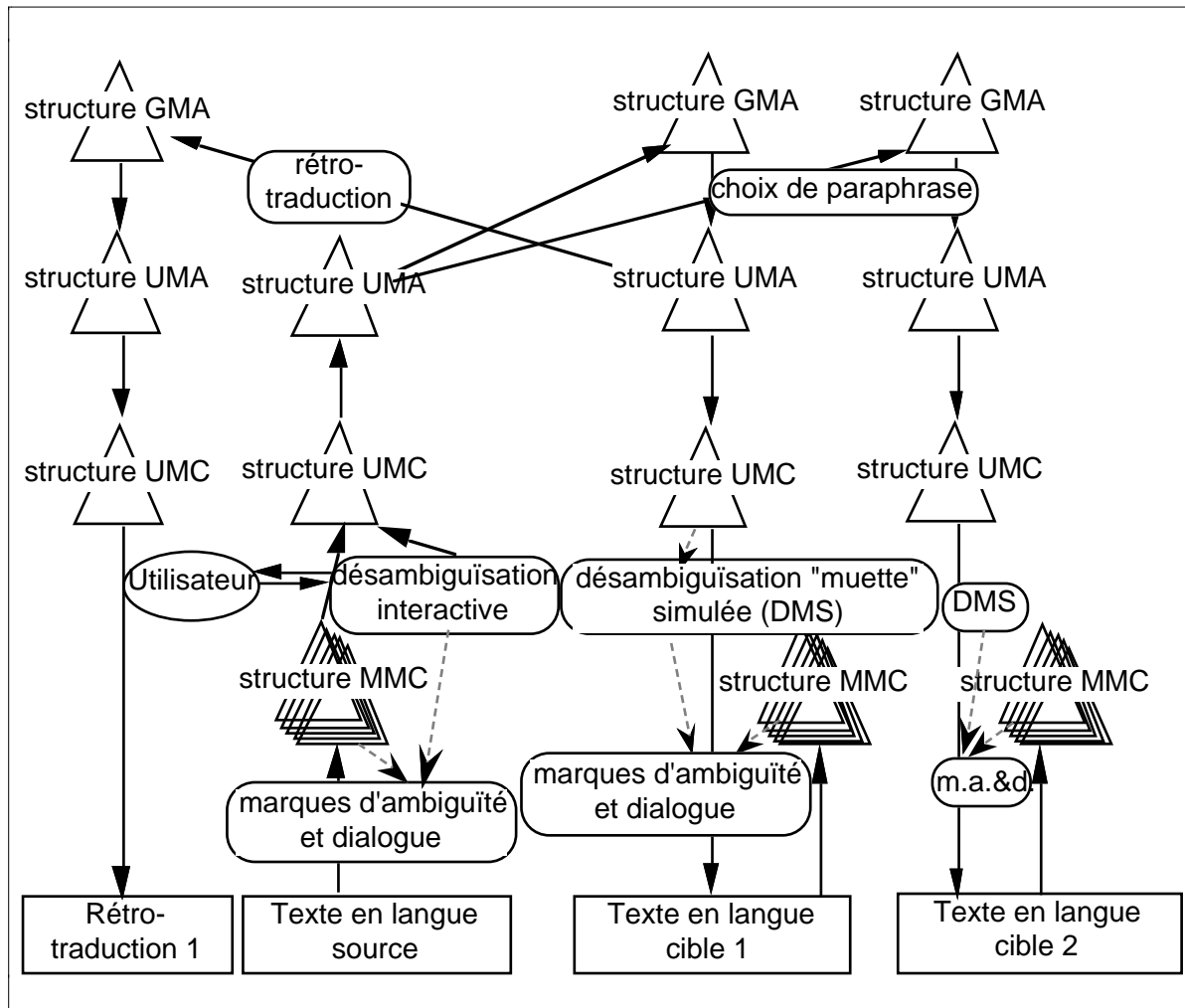
Enfin, les techniques et outils de génie logiciel modernes (essentiellement la programmation par objets), permettent de construire des systèmes complexes et interactifs bien plus rapidement et sûrement que par le passé.

Parmi les situations favorables à la TAFD avec entrée écrite, on peut mentionner :

- la traduction de volumes relativement faibles de documentation technique en plusieurs langues, typiquement 5.000 à 8.000 pages à distribuer sur un DON, par exemple dans les 9 langues officielles de l'UE, et peut-être dans d'autres aussi, comme le russe, l'arabe, le japonais, ou le chinois (1 → N, N≈10—20).
- la diffusion dans plusieurs langues d'informations sur la circulation, sur la météo, sur une manifestation internationale, sur une situation d'urgence... (1 → n, n≈2—5).
- l'échange télématique de notes et de documents de travail dans des projets internationaux (n ↔ n, n≈2—5).

Le schéma suivant illustre l'organisation des processus mis en œuvre dans un système de TAFD construit selon notre approche.

¹² ITS (Melby & al. 1980) demandait même *plusieurs* intervenants, un spécialiste du système pour l'analyse, et un bilingue pour chaque langue cible.



Voici les principes essentiels qui, pris ensemble, mènent à un "paradigme" qui nous semble nouveau, bien que presque tous ces principes aient déjà été proposés ou utilisés dans d'autres combinaisons.

- *Traitement distribué.* Le document est créé et désambiguïté interactivement sur un ordinateur personnel de milieu de gamme (un Macintosh dans notre maquette LIDIA-1.2), tandis que les traitements liés à la traduction automatique proprement dit sont effectués sur un serveur distant.
- *Désambiguïté interactive mais analyse automatique.* En effet, les analyseurs interactifs posent presque tous leurs questions dès qu'ils rencontrent un problème. L'ordre des questions suit le cours de l'algorithme, et n'obéit pas à des considérations d'ergonomie. D'autre part, l'utilisateur se rend vite compte que beaucoup de questions sont inutiles, car elles ne correspondent pas à plusieurs analyses complètes de l'énoncé.
- *Application à des documents structurés.* Dans LIDIA-1.2, les documents sont des piles HyperCard, mais bien d'autres formats sont possibles. On peut alors délimiter de petites unités textuelles, et typer celles qui sont à traduire.
- *Traitement asynchrone et non préemptif.* Une fois "relâchée" par l'auteur, une unité de traduction est envoyée au serveur de TA de façon transparente, revient après l'analyse sous une forme "multiple, multiniveau et concrète" (structure MMC), et annonce la présence d'ambiguïtés en faisant apparaître un bouton à côté d'elle.

Quand l'utilisateur le veut, il clique sur ce bouton, ce qui déclenche le dialogue de désambiguïté, après lequel l'unité de traduction, sous forme "unique, multiniveau et concrète" (structure UMC) retourne automatiquement vers le serveur, où elle est transformée en une forme "unique, multiniveau et abstraite" (structure UMA), puis traduite vers les langues

cibles désirées. Chaque traduction revient vers le micro de l'auteur, et est insérée dans le champ adéquat de la pile cible créée automatiquement par copie de la pile source.

- *Communication par courrier électronique avec le serveur de TA.* Cette technique, possible dans un contexte asynchrone, permet au serveur d'être situé n'importe où, et ne demande pas que le client et le serveur soient actifs en même temps.
- *Transfert multiniveau avec acceptions interlingues.* Nous avons ajouté aux niveaux lexicaux classiques de l'approche par transfert multiniveau de B. Vauquois ("occurrence" ou forme, "lemme" ou forme de citation, et "unité lexicale" ou famille dérivationnelle) un niveau d'acceptions interlingues, correspondant à l'union des sens des termes dans les langues considérées.
- *Système unique adaptable et approche par langage guidé.* Plutôt que de forcer les utilisateurs à utiliser un unique "langage contrôlé", nous proposons d'associer un "type de document" au document entier, un "genre de texte" aux fragments de la hiérarchie logique (éventuellement définie par une grammaire SGML, avec des attributs linguistiques), et un "style d'énoncé" à chaque fragment de base (phrase, titre, cellule de tableau...). Cette association, dite "typage textuel", doit être faite interactivement, en guidant l'utilisateur. Elle permet de limiter l'espace de recherche de l'analyseur et par suite le nombre de questions. De façon orthogonale, le dictionnaire, unique, peut être spécialisé par adaptation de poids ("préférences lexicales") portés par les arcs et les nœuds du réseau représentant la base lexicale.
- *Stratégie de désambiguïsation.* Nous avons développé un générateur de dialogues de désambiguïsation qui ne repose pas sur des traitements linguistiques trop sophistiqués, et peut tourner en temps réel sur un micro. Les règles de désambiguïsation ne sont pas intégrées à l'analyseur, mais exécutées sur son résultat, et organisées en fonction de la stratégie choisie. Une règle est une liste de couples (schéma, action). Les actions construisent les items des dialogues de désambiguïsation, de façon assez simple pour que des non spécialistes puissent facilement les comprendre (Blanchon 1994, Blanchon 1995).
- *Contrôle par rétrotraduction.* L'auteur, supposé monolingue, peut demander une

"rétrotraduction" pour contrôler ce que le système va produire dans une langue cible donnée. Pour effectuer une rétrotraduction, on ne part pas de la traduction elle-même, mais de la structure UMA produite lors de la génération. Cette structure est exactement la même que celle qui aurait été obtenue en réanalysant le texte cible, en désambiguïsant la structure MMC obtenue en faisant les choix correspondant à ceux faits pour la source, et en transformant la structure UMC résultante en une structure UMA par "abstraction".

- *Homogénéité des sources de connaissances lexicales.* En TAFD, le dictionnaire utilisé par le désambiguïseur interactif et les dictionnaires du serveur de TA doivent impérativement être homogènes. Sinon, les sens choisis ne correspondront pas aux sens traduits. Dans notre implémentation, ces dictionnaires sont construits à partir d'une BDLM unique, PARAX (Blanc 1993).

3.2 Dictionnaires : THAM → TAFD

Dans la majorité des situations adaptées à la TAFD, il faut un système de couverture lexicale et grammaticale très large. Sachant qu'on n'obtient de bons résultats que sur des langages restreints, on se heurte au problème de construire une base de connaissances linguistiques support du traitement de sous-langages multiples. Nous cherchons à le résoudre en séparant les aspects grammaticaux et lexicaux.

Laissons de côté ici les premiers¹³, et intéressons-nous au problème de la construction de BDLM pour la TAFD "pour tous". D'abord, la taille de ces bases sera bien plus grande que celle des dictionnaires de TA. En effet, un système de TAFD classique contient de 3.10^4 à 3.10^5 termes, en 2 langues¹⁴. Mais un système de TAFD visant le grand public et non restreint à un domaine particulier demandera de 3.10^5 à 3.10^6 termes, en plusieurs langues.

De plus, dans des situations fortement multilingues, l'approche par interlingua est séduisante. Sachant qu'un dictionnaire "conceptuel", est très difficile (et donc coûteux) à construire, comme les projets ATLAS (Fujitsu), PIVOT (NEC), EDR, et

¹³ voir (Boitet 1993), ou (Boitet 1994a), ou l'article de K. Grasson à cette conférence.

¹⁴ Le cas de METEO (3.10^3) est atypique, à cause de son domaine très restreint (Chandioux 1988).

CICC (ODA) l'ont montré, nous nous limitons, comme le projet ULTRA de la NMSU (Huang 1990), à un dictionnaire d'acceptions interlingues.

Il s'agit tout de même d'un travail plus lourd que la construction de BDLM pour des systèmes de TA classiques. C'est une raison de plus pour utiliser la synergie entre THAM et TA pour le faire. Nous pensons même qu'on ne pourra pas construire la BDLM d'un système de TAFD "pour tous" sans une "contribution lexicale généralisée" de traducteurs occasionnels bénéficiant de l'usage, gratuit par ailleurs, d'un outil de THAM distribué.

3.3 Mémoires de traduction structurées : TAFD → THAM

Une possibilité très intéressante apportée par la TAFD est de produire un document et toutes ses traductions dans un format "auto-explicatif", gardant la mémoire des ambiguïtés et du dialogue de désambiguïsation. On peut alors construire un visualiseur qui, sur demande, met en relief les parties ambiguës, et, à un deuxième niveau de détail, montre les différents sens possibles, et celui qu'il faut retenir.

Cette idée, motivée par la recherche en TAFD, est déjà très intéressante en contexte monolingue, par exemple pour produire des documents contenant leur "explication de texte", qu'il s'agisse de manuels techniques, de contrats, ou encore de textes normatifs.

Le schéma précédent montre comment, en TAFD, on pourrait produire une traduction en format auto-explicatif. Il faudrait disposer d'un analyseur de la langue cible, et du désambiguïseur interactif associé. On analyserait la traduction produite, et on ferait marcher le désambiguïseur en automatique, en choisissant, à chaque question, l'item qui sélectionne la structure (UMC) d'où est parti le générateur.

On pourrait alors structurer les mémoires de traduction, qui sont actuellement "plates", en format auto-explicatif. Ce serait un juste retour de la TA vers la THAM. Dans ce contexte, si on retrouve un fragment dans la mémoire phraséologique, en format auto-explicatif, on évite de refaire l'analyse, et on présélectionne les réponses aux questions correspondant à l'interprétation non ambiguë stockée, ce qui a toutes chances d'accélérer beaucoup la désambiguïsation interactive.

3.4 Télédéveloppement en TA et serveur lexical

Le réseau permet d'ores et déjà de construire des systèmes de TA, et donc aussi de TAFD, par "télédéveloppement". Par exemple, en utilisant le système CASH programmé par E. Blanc en HyperCard, on peut développer depuis un Macintosh des systèmes Ariane-G5, qui sont compilés et exécutés sur un serveur IBM-9221/130, les échanges se faisant ici encore par courriel.

Même si on produit les dictionnaires de TA et de THAM à partir d'une unique BDLM, on ne le fait pas tous les jours, et une dérive est inévitable. D'autre part, cette technique d'extraction totale produit des dictionnaires de TA et de THAM de taille maximale. Or, les flots de traduction n'ont jamais besoin de la totalité des informations lexicales, mais seulement d'une petite partie.

La disponibilité du réseau permettra de résoudre ces deux problèmes, en transformant la BDLM en "serveur lexical incrémental", communiquant courriel avec les divers composants utilisant de l'information lexicale. Il faut alors modifier légèrement le fonctionnement de ces composants.

Par exemple, avant d'appliquer une grammaire du mot inconnu, un analyseur morphologique enverra une demande à la BDLM. Si la réponse n'arrive pas assez vite, il traitera le mot comme un mot inconnu. Sinon, ou la prochaine fois que ce mot sera rencontré, il intégrera dans son dictionnaire interne les informations correspondantes.

À chaque consultation de la BDLM, il pourra aussi trouver des messages de mise à jour, dont il tiendra compte en fonction d'une stratégie ou d'une autre. Par exemple, si l'information sur un terme déjà présent dans le dictionnaire de l'analyseur est modifiée, on pourra la mettre à jour dans le dictionnaire interne, ou bien la supprimer, en attendant une prochaine occurrence pour la réinsérer, dans sa nouvelle version.

Enfin, si on dispose d'une communication très rapide avec le serveur, de type ATM par exemple, on pourra aussi envisager d'alléger encore la charge de la TAFD sur le poste client, en faisant tourner le désambiguïseur interactif sur le serveur.

On pourra ainsi passer de la TA du veilleur sur minitel de base à la TAFD de l'auteur sur les futurs ordinateurs de réseau.

Conclusion

La recherche d'une meilleure synergie entre TA et THAM, couplée à une utilisation plus intensive du réseau, nous semble ainsi pouvoir conduire à des progrès en TAO, d'ordre pratique ainsi que méthodologique.

Tout d'abord, en construisant des bases lexicales uniques pour la TA et la THAM, on améliorera la qualité de l'ensemble. Ensuite, développer un système de TA couplé avec un système de THAM tout en commençant à traduire par THAM, et ce éventuellement par télédéveloppement, lèvera les obstacles économiques actuels à l'introduction de la TA dans des situations fréquentes. Enfin, la construction de systèmes de TAFD "pour tous" sera rendue possible par couplage avec l'utilisation de systèmes de THAM, actuellement réservés aux professionnels, par des traducteurs occasionnels membres des communautés concernées, partageant les mémoires de traduction grâce au réseau.

Au niveau des dictionnaires, le progrès consistera non seulement à construire des bases lexicales multilingues (BDLM) servant de sources de connaissances à la TA comme à la THAM, mais aussi à modifier l'architecture des systèmes de TA pour que ces BDLM soient utilisées comme des serveurs lexicaux, de façon incrémentale.

Enfin, un retour de la TAFD vers la THAM pourrait consister à structurer les mémoires phraséologiques en format auto-explicatif. Cette forme permettrait sans doute aussi d'amorcer des systèmes de TA "par analogie" utilisant des distances entre formes arborescentes.

De façon plus générale, la construction et la mise à disposition par le réseau de ressources linguistiques de grande taille, créées en construisant et en utilisant des systèmes de TA et de THAM, ne pourra que bénéficier à la recherche en linguistique et en TALN, ainsi qu'à l'industrie de la langue.

Remerciements

La plupart des idées présentées plus haut ont été élaborées dans le cadre du GETA, et ont bénéficié de contributions trop nombreuses pour être énumérées ici. Que les contributeurs se reconnaissent et soient remerciés de leur apport ! Merci aussi au Pr. Moghrabi, qui m'a invité à rédiger une communication sur ce thème dans le cadre de cette conférence.

Références

- (Blanc 1993) *Visite guidée de PARAX, une base lexicale pentalingue par acceptions sous HyperCard*. GETA, IMAG, 30 p.
- (Blanchon 1994) *LIDIA-1 : Une première maquette vers la TA interactive "pour tous"*. Nouvelle thèse, UJF.
- (Blanchon 1995) *An Interactive Disambiguation Module for English Natural Language Utterances*. Proc. NLPRS'95, Seoul, 4-7 Dec. 1995, vol. 2/2, pp. 550-555, 6 p. (best paper award for the technical content and the presentation)
- (Boitet 1993) *La TAO comme technologie scientifique : le cas de la TA fondée sur le dialogue*. In "La traductique", Clas & Bouillon, ed., Presses de l'Université de Montréal, Montréal, pp. 109—148.
- (Boitet 1994a) *Dialogue-Based Machine Translation and Sub-Languages*. Proc. ICLA-94, 26-28 July 1994, USM, 14 p.
- (Boitet 1994b) *Perspectives of Machine-Aided Human Translation*. Proc. ICLA-94, 26-28 July 1994, USM, 4 p.
- (Boitet & Blanchon 1994a) *Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup*. Machine Translation, 9/2, pp. 99—132.
- (Boitet & Blanchon 1994b) *Promesses et problèmes de la "TAO pour tous" après LIDIA-1, une première maquette*. Langages, 116, pp. 20—47.
- (Boitet & Nédobekjine 1986a) *Toward integrated dictionaries for M(a)T: motivations and linguistic organization*. Proc. COLING-86, Bonn, IKS, pp. 423—428.
- (Boitet & Nédobekjine 1986b) *Vers des bases lexicales intégrées pour la T(a)O: motivations et organisation linguistique*. Proc. 8-ièmes Journées francophones d'informatique, janvier 1986, pp. 151—170.
- (Brown 1989) *Augmentation*. Machine Translation, 4, pp. 1299-1347.
- (Brown & Nirenburg 1990) *Human-Computer Interaction for Semantic Disambiguation*. Proc. COLING-90, Helsinki, 20-25 août 1990, Karlgren, ed., ACL, vol. 3/3, pp. 42-47.
- (Chandioux 1988) *10 ans de METEO (MD)*. In "Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires", Abbou, ed., Observatoire des Industries de la Langue (OFIL), Paris, mars 1988, pp. 169—173.
- (Chandler, Holden, Horsfall & al. 1987) *N-tran Final Report*. Alvey Project, 87/9, CCL/UMIST, Manchester.

- (Huang 1990) *A Machine Translation System for the Target Language Inexpert*. Proc. COLING-90, Helsinki, 20-25 Aug. 1990, Karlgren, ed., ACL, vol. 3/3, pp. 364-367.
- (Kay 1973) *The MIND system*. In "Courant Computer Science Symposium 8: Natural Language Processing", Rustin, ed., Algorithmics Press, Inc., New York, pp. 155-188.
- (Lehrberger & Bourbeau 1988) *Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation*. John Benjamins, 240 p.
- (Maruyama, Watanabe & Ogino 1990) *An Interactive Japanese Parser for Machine Translation*. Proc. COLING-90, Helsinki, 20-25/8/90, Karlgren, ed., ACL, vol. 2/3, pp. 257-262.
- (Melby, Smith & Perterson 1980) *ITS : An Interactive Translation System*. Proc. COLING-80, Tokyo, 30 septembre-4 octobre 1980, Nagao, ed., pp. 424-429.
- (Sadler 1989) *Working with analogical semantics : Disambiguation technics in DLT*. Witkam, ed., Distributed Language Translation (BSO/Research), Floris Publications, Dordrecht, Holland, 256 p.
- (Sérasset 1994a) *Approche œcuménique au problème du codage des structures linguistiques*. Proc. TALN-94, 7-8 avril 1994, pp. 109—118.
- (Sérasset 1994b) *An Interlingual Lexical Organization Based on Acceptions*. Proc. ICLA-94, 26-28 July 1994, USM, 12 p.
- (Sérasset 1994c) *Recent Trends of Electronic Dictionary Research and Development in Europe*. EDR, Japon, mars 1994, 88 p.
- (Sérasset 1994d) *SUBLIM, un système universel de bases lexicales multilingues; et NADIA, sa spécialisation aux bases lexicales interlingues par acceptions*. Nouvelle thèse, UJF (Grenoble 1).
- (Somers, Tsujii & Jones 1990) *Machine Translation without a source text*. Proc. COLING-90, 20-25 Aug. 1990, Karlgren, ed., ACL, vol. 3/3, pp. 271-276.
- (Tomita 1986) *Sentence Disambiguation by asking*. Computers and Translation, 1/1, pp. 39-51.
- (Wehrli 1991) *Pour une approche interactive au problème de la traduction automatique*. Proc. Colloque "L'environnement traductionnel. La station de travail du traducteur de l'an 2001", Mons, 25-27 avril 1991, AUPELF & UREF, pp. 59-68.
- (Wehrli 1992) *The IPS System*. Proc. COLING-92, Nantes, 23-28 July 1992, Boitet, ed., vol. 3/4, pp. 870-874.
- (Whitelock, Wood, Chandler & al. 1986) *Strategies for Interactive Machine translation : the experience and implications of the UMIST Japanese project*. Proc. COLING-86, Bonn, 25-29 août 1986, IKS, pp. 25-29.
- (Wood 1989) *Japanese for speakers of English: The UMIST/Sheffield Machine Translation Project*. In "Recent Developments and Applications of Natural Language Processing", Peckham, ed., Kogan Page Ltd, London, pp. 56-64.

-0-0-0-0-0-0-0-0-0-0-