LA TRADUCTION ASSISTÉE PAR ORDINATEUR (TAO) À GRENOBLE

TAMMIE DIANE LINGARD

La Traduction Assistée par Ordinateur (TAO) à Grenoble

par

Tammie Diane Lingard, B.A. B.Ed.

Thèse déposée pour satisfaire aux exigences en vue d'obtenir le grade de Maîtrise ès Arts.

Département d'études françaises et espagnoles
Memorial University of Newfoundland

1995

St. John's Newfoundland

ISBN 0-612-13918-2

Cette thèse a été examinée et approuvée par:

Examineur interne _____

Examineur interne _____

La Traduction Assistée par Ordinateur (TAO) à Grenoble

par Tammie Diane Lingard

Au cours de cette étude, ayant indiqué les grandes lignes des recherches préliminaires, nous nous attacherons aux idées directrices des Grenoblois.

Dans les années soixante, le Centre d'Études pour la Traduction Automatique (CETA) a construit un système de TAO russe-français, fondé sur le passage par un langage pivot hybride, de qualité suffisante pour permettre à un spécialiste de comprendre des textes russes scientifiques.

En 1970, le Groupe d'Étude pour la Traduction Automatique (GETA) a élaboré un système fournissant des traductions brutes qui puissent être révisées rapidement. Le générateur de systèmes développé, ARIANE-78, permet d'écrire des maquettes multilingues.

Dans les années quatre-vingt le GETA contribue vivement à des transferts de technologie vers l'industrie. En 1982, ARIANE-78 a été choisi pour la mise en oeuvre d'un prototype industriel dans le cadre du Projet National de TAO. ARIANE a contribué également à la définition de l'architecture du projet EUROTRA.

Le système ARIANE-78, remanié et étendu, a donné naissance à l'actuel ARIANE-G5. Pourtant, des faiblesses ont

été détectées: ces systèmes ne sont réellement efficaces que pour de gros flux de textes homogènes et seulement s'ils sont maintenus et développés par des équipes d'une même organisation.

Depuis 1989, des recherches en TAO du réviseur se poursuivent, mais l'orientation principale reste la TAO du rédacteur. Les recherches sont surtout axées sur une traduction de type personnelle. Le but principal du projet LIDIA est de mettre la TAO à la portée des rédacteurs (monolingues) de documentation technique. Une maquette limitée permet de simuler les fonctionnalités de traduction du poste de travail du rédacteur et traite de documentations multilingues hétérogènes. C'est l'équipe du projet NADIA qui cherche par la suite à formaliser et à implémenter des bases lexicales multilingues.

Le GETA se limite à la maintenance d'ARIANE-G5. Le système est employé par le projet LIDIA, par des partenaires industrielles, et dans le cadre du projet EUROLANG. L'équipe poursuit sa collaboration avec des partenaires commerciaux européens afin de créer une "trousse à outils" pour la TAO.

Pendant la période 1993 - 1995 les chercheurs comptent développer la maquette LIDIA-1 et réaliser des outils permettant le passage entre NADIA et les sources lexicales accessibles.

AVANT-PROPOS

Je tiens à remercier tout particulièrement mon directeur d'études, le Professeur John Hare, pour ses critiques constructives et bienveillantes et son encouragement pendant la rédaction de ce mémoire.

Je voudrais exprimer ma profonde gratitude au Groupe d'Étude pour la Traduction Automatique dont la coopération et les conseils me furent précieuses, ainsi que K. Levenbach (Linguiste - Documentaliste) qui m'a souvent aidée. Mes remerciements vont surtout à M. Christian Boitet, président recherche du GETA, qui a bien voulu m'aider dans la rédaction préliminaire.

Je reconnais de même l'immense service qui me fut rendu par les employés à la Bibliothèque de la Memorial University of Newfoundland.

Acceptez, je vous prie, mes très sincères remerciements pour votre contribution à mon travail.

TABLE DES MATIERES

LISTE DE FIGURES

ABRÉVIATIONS ET SIGLES

ADI	Agence pour le Développement de l'Informatique
ALPAC	Automatic Language Processing Advisory Committee
ATEF	Analyse de Textes en États Finis
AUPELF	Association des Universités Partiellement ou Entièrement de Langue Française
CALLIOPE	CALcul LInguistique OPErationnelle
CCL	Center for Computational Linguistics
CEE	Communauté Économique Européenne
CEDOCAR	CEntre de DOCumentation de l'ARmement
CELTA	Centre d'Études Linguistiques pour la Traduction Automatique (Nancy)
CERTAL	Centre d'Études et de la Recherche en Traduction Automatique (Paris)
CETA	Centre d'Études pour la Traduction Automatique
CETA	Contrôle ET Transductions d'Arborescences
CNRS	Centre National de la Recherche Scientifique
CRIN	Centre de Recherche en Informatique de Nancy
DGRT	Délégation Générale de la Recherche et de la Technologie
DIELI	Direction des Industries Électroniques et de l'Informatique
DRET	Direction des Recherches et Études Techniques
DRME	Direction de Recherche et Moyens d'Essais

EDR	Electronic Dictionary Research (Japan)
ESPRIT	European Strategic Program for Research and Development in Information Technology
EUROTRA	EUROpean TRAnslator
FAHQMT	Fully Automatic High Quality Machine Translation
FAHQT	Fully Automatic High Quality Translation
FAHQTUT	Fully Automatic High Quality Translation of Unrestricted Text
GDR	Groupement de Recherche
GETA	Groupe d'Étude pour la Traduction Automatique
IA	Intelligence Artificielle
IHM	Interface Homme\Machine
IMAG	Informatique et Mathématiques Appliquées de Grenoble
IMAG-LN	Informatique et Mathématiques Appliquées de Grenoble - section Langues Naturelles
ITEP	Ingénierie Technique et Publicitaire
LIDIA	Large Internationalisation des Documents par Interaction avec leurs Auteurs
LISP	LISt Processing Language
LSPL	Langage Spécialisé pour la Programmation Linguistique
MTS	Machine Translation Summit
NADIA	Une Nouvelle Approche de Dictionnaires Interlingues par Acceptions
NSF	National Science Foundation

PN-TAO	Projet National - Traduction Assistée par Ordinateur
PRC-CHM	Programme de Recherche Coordonnées - Communication Homme-Machine
PS	Personal System
SE	Systèmes Experts
SG2	Société de Gestion de la Société Générale
SITE	Sonovision ITEP TEchnologie
SMART	Système de recherche documentaire
SYGMCR	SYstème de Génération MORphologique
TA	Traduction Automatique
TAI	Traitement Avancé de l'Information
TALN	Traitement Automatique des Langues Naturelles
TAO	Traduction Assistée par Ordinateur
TAUM	Traduction Automatique Université Montréal
THAM	Traduction Humaine Assistée par la Machine
UJF	Université Joseph Fourier
UREF	Université de Réseaux d'Expression Française
VINITI	Institut de l'Union de Toutes les Républiques Socialistes Soviétiques de l'Informatique Scientifiques et Techniques
VM\CMS	Virtual Machine \ Conversational Monitor System
VMSP\CMS	Virtual Machine System Product \ Conversational Monitor System

1.0 INTRODUCTION

Les recherches menées en France dans le domaine de la traduction automatique ont commencé avec un certain retard par rapport aux pays pionniers, mais les chercheurs français ont bénéficié, dès le début, d'un climat favorable au développement de leurs propres programmes. Les études menées ailleurs ont stimulé l'élaboration de la traduction automatique.

Puisque nous avons l'intention d'analyser cette évolution, il convient de fournir quelques renseignements de base sur les recherches dans ce domaine avant l'éclosion des études françaises. En premier lieu, nous préciserons brièvement l'état d'avancement des travaux dans la traduction automatique en insistant sur les points forts et les lacunes de ces études. Cette démarche nous permettra de mieux cadrer l'apport français.

Une fois ces jalons posés, nous nous pencherons sur les recherches exécutées en France, en nous limitant à celles des Grenoblois. L'équipe, qui travaille sans relâche depuis plus de trente ans, s'est d'emblée écartée de la voie suivie par la plupart des équipes de recherche, plus soucieuses de répondre à des besoins toujours grandissants en traductions que de conduire des recherches informatiques et linguistiques de longue haleine. En examinant les principaux problèmes identifiés et abordés, les projets développés, et les voies

qui restent à explorer, nous ferons ressortir les contributions apportées par l'équipe de Grenoble.

Une étude préliminaire de ces travaux met en évidence l'étendue des difficultés qu'il faut surmonter pour obtenir un modèle pratique. Elles entraînent aussi des conséquences importantes dans la théorie de la traduction et dans la linguistique contemporaine. Néanmoins, jusqu'ici la documentation disponible au sujet de la traduction automatique traite soit d'un panorama de son développement dans le monde entier, soit d'un résumé d'un projet particulier. Nous situerons notre sujet par rapport à ce que d'autres chercheurs ont accompli (ou écarté) et par rapport à l'état d'avancement des connaissances internationales. En questionnant le thème de départ de cette façon, nous espérons établir une perspective valable sur la traduction automatique.

Nous retracerons donc l'évolution des projets à Grenoble; comment une trentaine d'années d'expérience a permis à un groupe de chercheurs de modifier les données de leurs problèmes. Ce travail de synthèse n'apporte aucune contribution à l'avancement de ces recherches mais éclaircira les difficultés croissantes dans la détermination de ce que l'homme peut attendre de la machine. Nous nous limitons à un exposé des grandes lignes dans le développement des systèmes de traduction automatique, en montrant pour chacun des modèles successifs les buts envisagés, la méthode utilisée, les

limites imposées par le système et l'utilité du modèle. De plus, nous nous bornerons aux systèmes qui ont comme but de fournir des textes "complètement traduits", sans examiner les systèmes "mi-automatisés" dont se servent les traducteurs humains, tels que les dictionnaires automatiques ou les banques de données terminologiques.

Cette nouvelle technologie recouvre un grand nombre d'applications dont la traduction automatique n'est qu'un seul exemple. Les systèmes de traduction automatique ont la particularité de permettre à l'utilisateur d'obtenir des textes d'arrivée. Toutefois, parmi ces systèmes, ceux qui sont opérationnels et qui semblent fournir des traductions d'une qualité acceptable n'obtiennent ces résultats, dans la plupart des cas, qu'au moyen d'une intervention humaine à différents stades de la traduction. L'expérience acquise à Grenoble montre que quel que soit le système envisagé, la traduction automatique doit être guidée ou corrigée à un moment ou à un autre par l'homme. Dans l'analyse des systèmes de traduction, nous adoptons les divisions conçues par les chercheurs Grenoblois.[1]

(i) Traduction du veilleur - sans révision, ni préédition

[1] Ch. Boitet. "Twelve Problems for Machine Translation". International Conference on Current Issues in Computational Linguistics, Penang, 12-14 juin 1991, 2.

ni postédition, destinée à produire des traductions "grossières", en grand volume et à bas coût, permettant de comprendre le sens du document original, en lecture rapide (donc avec le moins possible de marques ou symboles extratextuels, et une moindre importance de la grammaticalité).

(ii) Traduction du réviseur - visant à produire automatiquement des "premiers jets" dont la révision rapide par des professionnels fournirait la traduction à un rythme accéléré et à moindre coût. Les traductions sont subjectivement acceptables par les réviseurs, et révisables (grâce à une bonne grammaticalité, à des marques de choix lexicaux, à des avertissements en cas de doute sur la construction, ou sur telle ou telle propriété). Elle n'est envisageable que pour les textes homogènes extensifs, comme des manuels d'utilisation ou de maintenance et ne s'adresse qu'à des spécialistes au moins bilingues, et non à la plupart des rédacteurs qui sont unilingues.

(iii) Traduction du rédacteur - traduction de type personnelle pour rédacteur monolingue. L'on envisage de mettre la traduction à la portée des rédacteurs de documentation technique ou scientifique. Les systèmes traitent de nombreuses documentations multilingues trop hétérogènes pour être abordées par les systèmes de TAO

du réviseur.

(iv) Traduction du traducteur - La "Traduction Humaine Assistée par Machine" (THAM), selon laquelle l'utilisateur traduit au moyen d'un "poste de travail", adapté au traducteur-réviseur, peut offrir de bons résultats mais il s'agit ici de traduction assistée, et non plus automatique.

La matière ainsi délimitée a été traitée selon le processus suivant : nous divisons le développement de la traduction automatique à Grenoble non pas chronologiquement mais selon les projets de traduction automatique. Puisque les projets connaissent chacun leur propre évolution, cette division fera mieux ressortir la spécificité de chaque ensemble de recherches.

La première section traitera des renseignements de base en ce qui concerne l'historique de la traduction automatique. La deuxième portera sur la genèse des recherches entreprises à Grenoble. Dans la dernière section nous aborderons les projets contemporains. Évidemment il est fort difficile de faire des prédictions dans un domaine où les recherches évoluent non pas d'année en année, mais de mois en mois. Pourtant les horizons qui y sont révélés témoignent de l'évolution des conceptions principales des programmes actuels.

Tout aussi bien que dans la traductologie en général, la terminologie ne s'est pas encore stabilisée. Nous avons donc indiqué dans l'Annexe certains problèmes de délimitation sémantique et des définitions que nous avons adoptées dans cette étude.

2.0 CONTEXTE HISTORIQUE

2.1 PÉRIODE EXPÉRIMENTALE ET D'OPTIMISME

L'automatisation des processus de traduction n'est pas un rêve qui date d'hier. Cependant, s'il y avait des projets pour des machines à traduire dans les années trente[2], la traduction automatique est née avec la disponibilité du calculateur numérique. Le premier projet valable date de 1946, juste après l'avènement des premiers ordinateurs. Cette année-là Booth et Richens effectuèrent des expériences simples en Angleterre, surtout sur la réalisation des dictionnaires automatiques. L'automatisation répondait à une nécessité de l'époque car il fallait résoudre des problèmes de quantité et de vitesse dans la traduction. Andrew D.Booth suggéra à Warren Weaver, vice-président de la Fondation Rockefeller, que les calculatrices numériques pourraient servir à faciliter la tâche des traducteurs.[3] Ne comprenant pas qu'il est souvent

---

[2] M. Zarechnak. "The History of Machine Translation." Henisz-Dostert, B. et al. Machine Translation. The Hague: Mouton, 1979, 3-87.

[3] Booth était le directeur du laboratoire de calcul électronique à Birkbeck College, Londres. Selon Mounin, Booth voulait trouver de nouveaux marchés pour les calculatrices électroniques: "Lors d'une conversation avec Warren Weaver, en 1946, il (Booth) oriente l'intérêt de celui-ci... vers l'idée du dictionnaire automatique... Si l'on insiste ici, c'est moins par souci de rendre justice que par désir de bien marquer l'exact enchaînement des faits et des causes. Weaver ne parle pas de sa conversation de 1946 avec Booth, dans le fameux Mémorandum qui fait de lui le promoteur indiscuté de

indispensable de transformer la structure de la phrase, Weaver émet l'hypothèse qu'il suffira de remplacer les mots d'une phrase initiale par leur "équivalent" afin d'obtenir une traduction intelligible. Espérance prématurée, de nature à gêner la recherche, parce qu'elle encourageait une certaine

soulevé la vieille idée d'un "langage universel" intermédiaire entre les langues.[6] Les plans de Weaver, en partant du principe que le processus de la traduction ressemble au décodage des messages militaires et diplomatiques chiffrés, ainsi du ressort de la traduction automatique, ont suscité bien des recherches.

Dans les années cinquante, l'intérêt et l'appui financier ont été soutenus par des rêves de la traduction rapide de haute qualité de n'importe quel texte. Les premiers systèmes tentaient invariablement de produire des traductions "mot-à-mot", cherchant chacun des mots dans un dictionnaire bilingue, trouvant les équivalents dans la langue d'arrivée et fournissant le résultat dans le même ordre que le texte de départ. Nous aurons recours à la formule "première génération" pour désigner ce type de système. Évidemment, cette méthode était peu satisfaisante et bientôt on tentait de réarranger l'ordre des mots. C'est-à-dire, qu'afin de

---

[6] Selon Delavenay, Weaver souligne (1949) l'existence, "par-dessous les apparences diverses des langues, d'invariances statistiques... Ces invariances correspondraient à l'existence de caractères fondamentaux du cerveau humain et aux origines psycho-sociales communes du langage." Émile Delavenay. La machine à traduire. Paris: PUF, 1972, 36.

La recherche des "universaux" et des "invariants" sémantiques, reflétée dans les théories prochaines, rencontre une évolution déjà existant. Dans son Mémorandum, (W. Weaver. "Translation," 16). Weaver constate la possibilité de "descendre, de chaque langue, vers la base commune de la communication humaine, (ce qui serait) bien que non découvert jusqu'ici, le véritable langage universel."

traduire d'une manière acceptable, il fallait employer une analyse syntactique. De plus, dans cette première période de la traduction automatique, les limitations du matériel des ordinateurs et les insuffisances des langages de programmation constituaient des problèmes critiques qui sont toujours loin d'être résolus.

Pendant cette époque, en règle générale, la linguistique n'a pas eu d'impact sur la conception des machines à traduire. Les premiers travaux consacrés à la traduction automatique ne faisaient guère référence ni à la linguistique, ni aux linguistes. L'urgence de leurs efforts a poussé les chercheurs à vouloir se passer des développements dans les études linguistiques. La tradition de Bloomfield qui a dominé la linguistique américaine dans les années cinquante-soixante fixait l'attention sur des techniques descriptives et sur les problèmes de la phonologie et de la morphologie ; l'on ne s'intéressait guère ni à la syntaxe ni à la sémantique. Pourtant, certains chercheurs ont développé des méthodes d'analyse syntactique basées sur des fondations explicitement théoriques. Par exemple, Paul Garvin formula son "fulcrum method" qui décrivait des structures syntagmatiques indiquant des relations de dépendance entre ses constituants.[7]

---

[7] P.L. Garvin. On Machine Translation: Selected Papers. The Hague: Mouton, 1972.

> "Not only is the sentence as a whole assumed to have a fulcrum, but the various constituents of a sentence in turn are assumed to have their fulcra... Once the predicate is known, it allows a reasonable prediction concerning possible subjects and objects." [8]

Cette méthode a été adoptée dans un projet de traduction à l'université de Wayne State. Pourtant, dix années de travail (1959-1972) ont révélé les défauts du système. Même un programme très complexe était incapable de traduire des phrases russes comprenant plus d'un verbe à un mode fini.[9] Bar-Hillel, et, par la suite, Chomsky ont exploré la possibilité de représenter figurément les connexions syntactiques de la phrase au moyen des formules de logique symbolique. Ils démontrent ainsi la possibilité de représenter symboliquement des syntaxes. Cependant, la valeur de ce formalisme mathématique réside dans sa capacité descriptive et il reste à établir la possibilité de passer du symbolisme logique de la syntaxe d'une langue donnée, au symbolisme logique d'une autre langue. A ce moment-là Chomsky avait déjà démontré que les modèles syntactiques de cette période, en particulier la version de structure syntagmatique, étaient en principe insuffisants dans la représentation et la

---

[8] Garvin. "Syntax in Machine Translation", dans On Machine, 223-232.

[9] H.H. Josselson et al. Fourteenth (Final) Annual Report on Research in Computer-Aided Translation Russian-English, Wayne State University, avril 1972.

description de la syntaxe des langues "naturelles".[10]

L'utilité des travaux dans cette première période, étant donné la qualité moyenne des traductions obtenues, est sujette à discussion. L'optimisme un peu naïf du début quant aux possibilités de traduire mot-à-mot, s'est heurté au problème des flexions, puis à celui de la détermination des catégories grammaticales des mots. La véritable importance de la grammaire est devenu rapidement évidente puisqu'on ne pouvait traduire sans tenir compte de la grammaire de la langue d'origine. C'est le cas lorsque la machine, devant la phrase: "I am giving the cat meat" ne peut identifier le rôle grammatical de "cat" ou de "meat". La méthode d'analyse syntactique capable de fournir des solutions à ce genre de problème demandera encore des recherches considérables.

Néanmoins, les postulats présentés ont joué un rôle capital dans l'ouverture des voies de recherche qui ne sont pas encore complètement explorées. Non seulement l'expérience des linguistes est d'un grand profit pour la traduction automatique, mais les machines à traduire, elles aussi, ouvrent des perspectives valables qui se révèlent d'un grand profit pour la linguistique. Le fait que la traduction automatique peut beaucoup contribuer à l'étude du langage est déjà évident en 1956 : "Une lumière nouvelle est ainsi versée

---

[10] N. Chomsky. Syntactic Structures. The Hague: Mouton, 1957.

sur maints phénomènes du langage et de la pensée. Le langage sera sans nul doute le premier bénéficiaire de ce progrès nouveau des connaissances." [11]

Le développement des premiers ordinateurs étant surtout dû aux chercheurs américains, c'est aux États-unis que débutèrent les travaux sur la traduction automatique. Les recherches dans ce domaine ont commencé à l'université de Washington, à l'UCLA, et à MIT. Au Massachusetts Institute of Technology, le logicien israélien Bar-Hillel est le premier chercheur à temps plein dans ce domaine et en particulier dans l'étude du langage aux fins de traduction automatique. Avec l'aide financière de la fondation Rockefeller, Bar-Hillel assiste dans l'organisation du premier congrès de linguistes et d'électroniciens. Consacrée aux problèmes de la traduction mécanique, cette réunion a eu lieu au printemps de 1952. Une des conclusions principales était que "la recherche nécessaire... était en premier lieu linguistique".[12] On y a aussi évoqué le perfectionnement d'un système de traduction mot-à-mot dans un avenir proche, donc le besoin d'une continuation des recherches et le projet de construire une machine expérimentale, capable d'une démonstration propre à

---

[11] E. Cary. "Mécanismes et traduction." Babel. 2:3 (1956): 104.

[12] Victor Yngve. "Report." Proceedings of the VIII International Congress of Linguists. Oslo: Oslo UP, 1958, 511.

convaincre le public, en même temps que les bailleurs de fonds.

En 1954 l'équipe de recherche à l'université de Georgetown a fait cette démonstration publique dans le dessein de démontrer la faisabilité technique de la traduction automatisée. Avec un vocabulaire de seulement 250 mots russes, six règles de grammaire et des phrases russes sélectionnées soigneusement, le système démontré n'avait aucune valeur scientifique. Le projet risquait de convaincre le grand public que les problèmes de la traduction automatique étaient déjà résolus mais a quand même réussi à attirer l'attention sur la traduction automatique comme entreprise réalisable.[13] En 1954 également, le MIT publie le premier numéro de la revue Mechanical Translation (M.T), dirigé par William Locke et Victor Yngve.[14] L'année suivante le premier ouvrage consacré à ce sujet, Machine Translation of Languages est publié sous la direction de Locke et de Booth.[15]

---

[13] Malgré son optimisme, Delavenay reconnaît les faiblesses des systèmes actuelles: "Le succès de cette expérience de traduction automatique de phrases entières, sans pré-édition ni révision, força l'attention et fit constater assez largement les progrès réalisés dans la recherche... L'expérience soulignait d'autre part l'importance des problèmes linguistiques..." Delavenay. La machine, 47.

[14] Mechanical Translation, devoted to the Translation of Languages with the Aid of Machines, Massachusetts Institute of Technology, 1954.

[15] William Locke et al. Machine Translation of Languages. New York: MIT UP et Wiley, 1955.

Entre les années cinquante et soixante l'appréhension aux États-Unis en ce qui concerne les avances russes dans les sciences et dans la technologie a incité l'appui massif des systèmes à l'essai dans la traduction russe-anglais. L'annonce du lancement réussi par l'Union soviétique d'un premier satellite Spoutnik, le 4 octobre 1957, provoqua une vive réaction des Américains qui mobilisèrent leurs efforts afin d'être les premiers à mettre un homme sur la lune. Pour rattraper et dépasser les équipes soviétiques, il leur fallait connaître non seulement les résultats des Russes, mais aussi l'orientation de leur technologie. A la suite du choc du premier Spoutnik, la traduction automatique était devenue un domaine de recherche prioritaire. Pendant la prochaine décennie, les adeptes subvenaient aux besoins de la recherche aux États-Unis sur une grande échelle - à dix-sept institutions moyennant des subventions de vingt millions de dollars.[16] Malgré l'état peu avancé des ordinateurs et des théories compréhensives du language, certains prétendaient que dans quelques années des systèmes capables de traduire n'importe quel texte pouvaient être construits. Les expériences menées à l'université de Georgetown ont encouragé les intéressés à croire que la traduction par ordinateur avait

---

[16] A.H. Roberts et M. Zarechnak. "Mechanical Translation." Current Trends in Linguistics, vol. 12: Linguistics and Adjacent Arts and Sciences, pt.4. The Hague: Mouton, 1974, 2825-2868.

été, en principe, effectuée.

La stratégie générale employée dans les systèmes de cette période jusqu'au milieu des années soixante a été l'approche de traduction directe. Les systèmes ont été conçus expressément en vue d'une seule paire de langues, presque toujours, à cette époque, pour le russe comme langue de départ et l'anglais comme langue d'arrivée. Un exemple typique est le système de l'université de Georgetown, qui s'est révélé un des exemples les plus fructueux de l'approche directe.[17] En 1964 des systèmes russe-anglais ont été livrés à l'"U.S. Atomic Energy Commission" et à Euratom en Italie. L'équipe de recherche de Georgetown a adopté ce que Garvin a appelé plus tard l'approche "brute force" [18] Il importe pour un couple de langues déterminé, de trouver des règles de transfert au niveau des expressions dans chacune de ces deux langues. Un programme aurait été rédigé à l'intention d'un seul texte en particulier, mis à l'essai sur un autre texte, modifié et amélioré, mis à l'essai sur un plus grand texte, modifié de nouveau, etc. Il s'ensuit que les programmes étaient monolithiques, complexes et sans aucune séparation entre les parties qui devaient analyser les textes de départ et celles

---

[17] R.R. Macdonald. Georgetown University Machine Translation Project. General Report 1952-1963. Washington, D.C., 1963.

[18] Garvin. On Machine.

qui devaient fournir les textes d'arrivée. L'analyse syntactique était rudimentaire. Par conséquent, il devenait de plus en plus difficile d'apporter des modifications au système. L'idée qu'il valait mieux choisir une seule paire de langues au départ, traduites dans un seul sens a été au coeur des travaux théoriques des machines de la première génération. Néanmoins, cette concentration des efforts sur une seule paire de langues servait à fournir des conceptions nouvelles de la notion de stylistique comparée.[19]

En 1955, après la démonstration de l'université de Georgetown, des savants de l'Union soviétique se sont mis à la recherche dans le domaine de la traduction. Des chercheurs appartenant à un nombre assez grand d'organisations, surtout à l'Académie des Sciences de l'URSS, ont entrepris les premières démarches. Aux années soixante il existe à Moscou un groupe dont les membres appartiennent à divers instituts et qui s'occupent de linguistique formalisée dans le but de réaliser la traduction automatique. Au sein de ce groupe, A. Zholkovskij et I. Mel'chuk ont présenté leur "modèle sens-texte". En mai 1957 un important congrès scientifique et

---

[19] Voir deux ouvrages importants dans ce domaine:
J-P. Vinay et J. Darbelnet. Stylistique comparée du français et de l'anglais: Méthode de traduction. Paris: Didier, 1958.
Jacqueline Guillemin-Flescher. Syntaxe comparée du français et de l'anglais: Problèmes de traduction. Paris: Ophrys, 1981.

technique sur les "Problèmes du Développement de la Construction des Machines à Information" avait accueilli beaucoup de communications dans lesquelles on constatait la nécessité de faire de cette matière une science exacte. Pendant longtemps les Russes exposaient des études théoriques de langue, sans aucun rapport avec la tâche de concevoir des systèmes pratiques. Après cette insistance sur la base théorique de la traduction automatique, les chercheurs reconnaissent la création éventuelle d'un système pratique dans un proche avenir et ils soutiennent le travail expérimental. En 1975 le Congrès international sur la traduction automatique à Moscou fait état de la reprise des travaux pratiques des linguistes.[20]

Au Japon, les études ont débuté en 1955 à l'université de Kyushu, dirigées par les professeurs Toshihiko Kurihara et Tsuneo Tamachi. Tamachi a été accordé une subvention gouvernementale pour aider sa recherche et assister dans la construction d'une machine à traduire, baptisée la KT-1. A ce projet ambitieux, s'ajoutent des études au laboratoire électrotechnique de l'Institut industriel du Ministry of International Trade and Industry (MITI), sous la surveillance du chef d'électronique, Hiroshi Wada. En 1959 le premier système de traduction automatique anglais-japonais fut

---

[20] Zarechnak. "The History", 75.

disponible au Japon.

C'est aussi en 1955 que la recherche a commencé en Europe. En Angleterre, les efforts les plus considérables ont été réalisés en collaboration avec les chercheurs américains au Cambridge Language Research Unit. Booth, Brandwood et Cleave publient en 1958 leur Mechanical Resolution of Linguistic Problems, tandis que Richens procède à l'étude de la possibilité d'une langue universelle de type algébrique.

En Italie, un postulat non-linguistique a été proposé par Silvio Ceccato, porte-parole de l'École Opérationnelle Italienne. Au lieu de se fier à la mécanisation du dictionnaire bilingue, typique dans les systèmes de la première génération, Ceccato estime qu'il faut étudier la nature de la pensée et les opérations mentales de l'homme. Une telle ambition implique la construction complète de l'activité pensante des hommes et, en effet, d'une théorie entière du langage: "it perhaps constitutes a step along the path that will eventually lead us to a mechanical translation identical with human translation..." [21]

En France, E. Delavenay est le fondateur d'un groupe de courte durée, l'Association pour la Traduction Automatique et la Linguistique Appliquée (ATALA). Le Professeur Vauquois,

---

[21] Silvio Ceccato. "Operational Linguistics and Translation." Linguistic Analysis and Programming for Mechanical Translation. Milan: University of Milan Italy, Giangiacomo Feltrinelli Editore, 1960, 48.

président de cette équipe française, dirigera le Centre d'Études pour la Traduction Automatique (CETA). A la direction de l'Armement du Ministre des Armées, le sujet de la traduction automatique n'était plus tout neuf lors de la création du CETA et le temps des explorateurs pionniers était déjà un peu dépassé. En 1960 le CETA, avec une section à Paris et une à Grenoble, passera sous le patronage du Centre Nationale de la Recherche Scientifique (CNRS). A cet institut de recherche, situé à l'université de Grenoble, on travaille encore aujourd'hui sous les auspices du CNRS. Le système de traduction au CETA de 1963 à 1969 est un système de traduction multilingue, destiné à traiter des textes scientifiques ou techniques, présentés avec une pré-édition réduite. Même après le rapport ALPAC, le CNRS ne cesse de poursuivre ces études. Particulièrement digne d'attention est le travail de Bernard Vauquois, les efforts duquel se sont prolongés jusqu'à sa mort en 1985.

Au milieu des années cinquante, les recherches dans ce domaine sont, en effet, internationales. En 1956, le MIT réunit la première conférence internationale à laquelle assiste une trentaine de spécialistes Canadiens, Anglais et Américains. Un document valable sur l'activité soviétique a été communiqué par le président de l'Académie des Sciences de l'URSS, D.J. Panov. Beaucoup de recherches se révèlent axées sur les mêmes problèmes et, assez souvent, les chercheurs,

sans s'en rendre compte, suivent pas à pas le même chemin. L'accord est unanime que la traduction automatique est possible et qu'il faut coordonner les recherches et les efforts de tous les adeptes.

2.2 INFLUENCE DES ÉTUDES LINGUISTIQUES

Vers 1959, et à un rythme accéléré dans les années suivantes, une conception nouvelle de la traduction automatique s'est dégagée, s'appuyant à la fois sur de nouvelles théories linguistiques, sur l'étude formelle des langues et sur la théorie des automates. Un des nouveaux objectifs s'explique par la notion de modèles "texte-signification" et "signification-texte". Le problème se pose de déterminer une notation au moyen de laquelle la signification des textes est enregistrée indépendamment du "langage de surface". Ce problème a été abordé par les linguistes soviétiques A. Jholkovskij et I. Mel'chuk. Leur système suit le principe de langue-intermédiaire universelle, construite sur la base de sémantique logique, repris dans les doctrines du CETA.[22] L'utilité d'un tel modèle, comme nous

---

[22] La particularité essentielle de la méthode de deux chercheurs russes, Mel'chuk et Zholkovskij, est le caractère sémantique de la synthèse. D'après ces auteurs, la synthèse "suppose une traduction prenant appui sur la signification ou "sens", par un processus analogue à ce que fait l'homme lorsqu'il traduit un texte : il commence par comprendre le texte..." I.A. Zholkovskij et I.A. Mel'chuk. "Sur la synthèse sémantique." T.A. Informations. 2 (1970): 2.

le verrons, est évidente dès qu'on cherche à construire des systèmes multilingues.

2.3 LE RAPPORT ALPAC

Les expériences dont nous avons fait état ne constituent pas un répertoire exhaustif des travaux de cette période. Cependant, ces expériences illustrent les démarches qui ont prévalu pendant cette décennie.[23] Au cours de ce rappel historique on se rend compte de l'état de la recherche dans plusieurs pays du monde dans les années cinquante, et de l'apparition d'un nombre de systèmes modestes de traduction automatique. La dissémination des revendications exagérées et les attentes trop optimistes, liées à la croyance à l'apparition prochaine des machines capables de traduire aussi bien que l'homme étaient normales. Au début des années 60, après le succès des premières expériences, certains ont cru à la possibilité imminente de traduction de haute qualité complètement automatisée. Pourtant, cinq années après la démonstration notoire de Georgetown aucun système n'avait été construit. Les adeptes ont envisagé la situation avec

---

[23] En 1961 Léon Dostert, Directeur du Laboratoire de l'université de Georgetown, récapitule la méthode des systèmes de la première génération: "La traduction est une affaire de mémoire (storage) et d'intelligence; la mémoire, c'est le dictionnaire ; l'intelligence c'est le programme". Bernard Vauquois. La traduction automatique à Grenoble. Les Documents de Linguistique Quantitative 24. Paris: Dunod, 1975, 29.

optimisme mais ont fini par se désillusionner : la traduction automatique a été sérieusement attaquée aux États-Unis en 1966-67. Le gouvernement américain a chargé un comité d'évaluer les recherches dans ce domaine. En 1964, la National Science Foundation a établi le "Automatic Language Processing Advisory Committee" (ALPAC) à l'instigation des promoteurs de la traduction automatique. Le célèbre rapport ALPAC est rendu public en 1966. La conclusion principale, souvent citée dans la documentation : "... we do not have useful machine trans-lation. Furthermore, there is no immediate or predictable prospect of useful machine translation." [24] Le rapport a ainsi rendu compte de l'état peu prometteur des recherches : la T.A. était plus lente, moins fidèle et coûtait deux fois plus cher que la traduction humaine.[25] Aux États-Unis les subventions furent supprimées et presque toutes les recherches cessèrent. Dans le rapport on insiste sur le fait que la linguistique doit être encouragée en tant que science ; on recommande la poursuite de

---

[24] ALPAC Language and Machines : Report of the Automated Language Processing Advisory Committee. Washington DC:National Academy of Sciences National Research Council, 1966, 32.

[25] B. Vauquois souligne le fait que, dans l'ensemble, le rapport ALPAC est très contestable: "il comporte de graves défauts: inexactitudes dans le recensement des faits, erreurs dans l'estimation de l'utilité de la T.A. (estimation subjective contredite par l'expérience), ignorance volontaire ou non des travaux en cours depuis 6 ans..." Vauquois. La traduction, 150.

trois voies de recherche :

(i) l'amélioration des techniques de traduction humaine ;

(ii) la linguistique computationelle ;

(iii) l'affirmation de la notion que la linguistique est une discipline scientifique et que l'on devrait étudier le langage lui-même, mettant de côté l'aspect pratique de telles recherches.

Il n'est pas difficile de comprendre pourquoi les auteurs du rapport ont tiré de telles conclusions. Les systèmes qui avaient été évalués étaient ceux qui avaient été conçus dans les années cinquante, quand les seuls modèles formels de la langue étaient ceux des analystes de la cryptographie comme l'avait proposé Weaver. Pendant sa deuxième décennie, dans les années soixante, la désillusion est entrée sur la scène : le nombre et la difficulté des problèmes linguistiques sont devenus de plus en plus évidents. L'on s'est bientôt rendu compte que le processus de la traduction n'était pas si prêt à l'automatisation que l'on ne l'avait pensé.

Le rapport ALPAC a mis fin à peu près partout aux recherches en traduction automatique mais il restait un côté positif. Le comité a recommandé le transfert de fonds de la recherche en traduction automatique au développement en linguistique computationelle. De cette façon la linguistique

avait la possibilité de rattraper les prétentions des programmes proposés. Malgré la publicité négative et le découragement financier à la suite d'ALPAC, dès le début des années soixante-dix on réclamait de nouveau la faisabilité d'une deuxième génération de systèmes de traduction automatique. Il s'agissait des propositions basées sur les leçons tirées des erreurs des années cinquante et du début des années soixante, de nouvelles idées glanées au cours de recherches parallèles dans les domaines de l'intelligence artificielle, de la linguistique computationelle et de la recherche documentaire, et d'une base théorique beaucoup plus solide.

Une des grandes leçons tirées des échecs des années soixante est que les parties du système décrivant le dictionnaire et le modèle linguistique utilisés auraient dû être clairement séparées de la partie logicielle où sont inscrits les algorithmes d'analyse, de transfert et de génération. Ceux de la première génération ne présentaient pas en effet une séparation aussi nette entre les logiciels et les données linguistiques. Ce sont des systèmes, comme SYSTRAN par exemple, qui ne s'inspirent pas des études linguistiques et dont il est difficile d'améliorer la qualité de la traduction. En revanche, ceux de la deuxième génération permettent d'améliorer les grammaires ou le modèle linguistique indépendamment de la partie logicielle du

système.

Tout en considérant la période de 1950 à 1965 comme le premier temps de la recherche dans cette sphère, l'on peut désigner la décennie de 1966 à 1975 "l'âge des ténèbres" dans l'histoire de la traduction automatique. A la suite du rapport ALPAC (1966), simplement par manque de subventions, il a été quasiment impossible de poursuivre des recherches dans ce domaine aux États-Unis. La périodicité de la revue M.T. devient très irrégulière et Panov juge que "la recherche est entrée dans une phase silencieuse, de travail et d'amélioration patiente des résultats déjà obtenus." [26] Inévitablement, la situation aux États-Unis s'est répercutée sur l'étranger, et l'on a été en perte de vitesse, premièrement en Angleterre, ensuite au Japon et dans le reste de l'Europe. Si l'on était conscient de l'impuissance des techniques syntactiques, il n'en résultait que peu de développements dans les méthodes sémantiques.

2.4 REGAIN D'ENTHOUSIASME

L'impact d'ALPAC a été moins sévère sur les pays où l'importance politique de la langue ne permet pas

---

[26] Mounin. La machine, 26.

traditionnellement la suffisance linguistique du monde anglais. Dans une revue d'avancement des théories et des programmes développés il ne faut pas délaisser le travail qui a été effectué, particulièrement dans l'Ouest, de 1965 jusqu'à la fin des années soixante-dix. L'effort de recherche et de développement est davantage soutenu en France ainsi qu'au Japon où les problèmes de traduction sont devenus importants pour des raisons économiques et politiques.

Les réclamations accrues vis-à-vis de la traduction automatique promues par le succès de ces systèmes, considérées conjointement avec des facteurs importants extérieurs, allaient aboutir à un regain d'enthousiasme pour la traduction automatique vers 1975, beaucoup plus tôt que l'on n'aurait pu l'envisager avant, étant donné la parution du rapport ALPAC. Pourquoi y a-t-il eu un redoublement d'efforts ? D'abord, par suite des recommandations du rapport ALPAC, plusieurs chercheurs travaillaient depuis le milieu des années soixante dans les sphères connexes de l'intelligence artificielle, de la linguistique computationelle et de la recherche documentaire ainsi que sur le problème de la "compréhension" des langues naturelles. Donc, au début des années soixante-dix, un nombre assez important d'idées nouvelles à propos du traitement des informations sémantiques et pragmatiques à l'intérieur de l'ordinateur se sont présentées.

Deuxièmement, l'importance politique croissante des

langues, doublée du coût toujours en hausse des traducteurs humains, commençait à occasionner une accumulation grave de textes à traduire. Cela était surtout le cas pour les pays comme le Canada et les organisations comme la Commission des communautés européennes, où le principe du bilinguisme ou du multilinguisme collectif fut établi constitutionnellement. Le développement le plus important, du moins pour l'Europe, s'est produit en février 1978, lorsque la commission européenne a réuni un comité de spécialistes de tous les pays membres de la commission afin de discuter du projet d'un système commun de traduction automatique. Ce système refléterait les avances des trois décennies précédentes et serait vraiment une image de l'état des connaissances et de la technique. Comme nous le verrons, EUROTRA deviendrait le projet officiel de la communauté en 1983.

Pendant cette période ont apparu des systèmes plus avancés en linguistique fondés sur des approches "indirectes". La recherche a également continué sur des systèmes de traduction "directe". Deux systèmes directes sont devenus pleinement opérationnels : LOGOS, conçu pour la traduction des manuels d'avion de l'anglais vers le vietnamien, est un système qui était en train d'être adapté à la traduction anglais-français, et espagnol-allemand.

Le système "direct" le plus reconnu est SYSTRAN. Conçu initialement pour la traduction russe-anglais et employé par

l'armée de l'air américaine depuis 1970, SYSTRAN a été livré à la Commission des communautés européennes (1976). A Bruxelles, au siège des Communautés européennes, tous les textes officiels sont traduits dans chacune des sept langues de l'alliance (allemand, anglais, danois, français, grec, italien et néerlandais), toutes des langues officielles, ayant la même valeur légale. Sans traduction, la Communauté ne fonctionnerait point. SYSTRAN a été choisi de préférence à d'autres systèmes opérationnels car il avait été employé avec succès par l'armée de l'air américaine pour la traduction russe-anglais et s'était trouvé en cours de développement pour l'anglais-français, un couple de langues essentiel pour la Communauté. On peut considérer SYSTRAN essentiellement comme un descendant fort amélioré du système "direct translation" de l'université de Georgetown. Mais si les améliorations en computation sont considérables, l'analyse linguistique est restée peu modifiée. Les avancements les plus importants sont l'aspect "modulaire" de sa programmation, permettant la modification de n'importe quelle partie du processus sans courir le risque de détériorer l'efficacité d'ensemble, et la stricte séparation entre les données linguistiques et les procédés opérationnels. Ainsi, ce système évite un grand nombre des complexités insolubles du système monolithique de l'université de Georgetown.

Par contre, les systèmes qui adoptent l'approche

"indirecte" ont bénéficié de l'apport des théories linguistiques. La possibilité de traduire au moyen d'un langage intermédiaire "universel" avait été suggérée par Weaver dans son mémorandum,[27] mais il fallait attendre jusqu'aux années soixante que la linguistique fournisse des modèles pratiques. L'approche du "langage pivot" a attiré deux équipes de recherche, au début des années soixante, à l'université du Texas et à l'université de Grenoble. A Grenoble, le système de traduction automatique du CETA a été fondé sur le principe du langage pivot, expression intermédiaire invariante dans le processus de la traduction. Selon Veillon :

> Bien entendu, l'invariant de toute bonne traduction doit être le sens. L'expression qui est utilisée ici est loin d'être de nature sémantique, mais doit être considérée comme une structure syntaxique profonde commune à un groupe de langues naturelles.[28]

Dans une étude (1960) Bar-Hillel ne doutait pas qu'à l'aide de la théorie linguistique l'on pût améliorer considérablement les méthodes de l'analyse syntactique. Cependant, il a exprimé la conviction selon laquelle les problèmes sémantiques

---

[27] Weaver, 15-23.

[28] G. Veillon. "Description du langage pivot du système de traduction automatique du CETA." T.A. Informations. 1 (1968): 8.

ne pourraient jamais être complètement résolus.[29] L'invariant idéal serait, selon certains, un langage pivot dans lequel on pourrait formuler tout texte et à partir duquel on pourrait transformer cette formulation pour restituer un texte équivalent dans une autre langue. Mais cet idéal reste inaccessible et selon les chercheurs du CETA doit être limité à un objectif bien plus modeste. Des essais d'analyse syntaxiques montrent que le logiciel du CETA n'a pas la fiabilité voulue et, indépendamment de toute considération d'ordre économique, ne peut servir tel quel de manière satisfaisante. En 1971 le CETA deviendra le Groupe d'Étude pour la Traduction Automatique (GETA) et l'ensemble aura plus d'influence après l'interruption des recherches américaines à la suite du rapport ALPAC.

---

[29] Selon lui la traduction automatique demanderait encore de longues études: "... FAQT is out of the question for the foreseeable future because of the existence of a large number of sentences the determination of whose meaning, unambiguous for a human reader, is beyond the reach of machines..." Yehoshua Bar-Hillel. "A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation." Advances in Computers. 1 (1960): 161.

3.0 LE CETA ET LA TA DU VEILLEUR

3.1 CRÉATION DU CETA

Ayant esquissé le contexte historique des travaux en cours à Grenoble, nous retracerons les travaux effectués par le Centre d'Études pour la Traduction Automatique pendant ses dix années d'existence (1961-1971).

Le CETA avait été créé en 1961 sous la direction de M. Sestier. Ensuite on le réorganisa en CETA-P, à Paris, et CETA-G, à Grenoble. Peu après, M. Sestier, convaincu que la traduction automatique présentait des difficultés insurmontables, se retira, dissolvant du même coup le CETA-P. Ce laboratoire du CNRS, autrefois partagé en deux sections, l'une à Paris, l'autre à Grenoble, est ainsi définitivement implanté à Grenoble depuis 1963. Lors de sa création en 1960, M. Emile Delavenay est le fondateur de l'Association pour la Traitement Automatique des Langues (ATALA). Il jouait une rôle important dans les premiers travaux et dans la création du CETA et d'autres laboratoires dans ce domaine au début des années soixante. L'Association devient le point de rencontre des adeptes de la traduction automatique et du traitement automatique des langues naturelles. Elle publie la revue T.A. Informations et organise annuellement des conférences à Paris auxquelles participent chercheurs français et étrangers.

Depuis l'origine, les ressources du CETA proviennent environ pour 75% du CNRS et pour 25% de la DRME.[30] C'est l'époque pionnière, où sous la direction de Bernard Vauquois, le premier système de "deuxième génération" fut construit. Persuadés que la traduction automatique était réalisable, les chercheurs se sont mêlés à la fois à des études linguistiques et au développement de la programmation automatique. L'objectif consiste à réaliser un système qui permettra la traduction de plusieurs langues vers le français.[31] Vauquois continue à la tête du CETA où, on le verra, des progrès considérables sont effectués. Des projets de traduction allemand-français, japonais-français et, sur une masse de données importante, de traduction russe-français ont été réalisés.

3.2 LANGAGE PIVOT

A cette époque la plupart des chercheurs, quelle que soit leur intention, envisagent des traductions sur un couple de

---

[30] Direction de Recherche et Moyens d'Essais (du Ministre de la Défense)

[31] Selon le CETA, la traduction automatique signifie "les activités de recherche qui visent à faire traduire des textes par ordinateur sans autre intervention humaine qu'une pré-édition réduite à des indications typographiques, des indications de "hors-textes" (tableaux, figures,...) et de formules, et, bien sûr, une révision par un humain compétent." Bernard Vauquois. Bernard Vauquois et la TAO, vingt-cinq ans de Traduction Automatique : ANALECTES. Boitet Ch., éd. Grenoble, Champollion & GETA, 1988, 506.

langues dans un seul sens, leur analyse de la langue de départ est donc strictement en fonction de la langue cible. Il n'y a ni structures arborescentes, ni règles de grammaire. Cette démarche, de traduction directe, ne s'appuie ni sur des théories linguistiques, ni sur des théories de langages formels, pour la bonne raison que ces théories n'ont été formulées qu'environ dix ans après le début des travaux sur la TA (1949). De l'autre côté, ceux qui se penchent sur la traduction multilingue considèrent la notion de langue intermédiaire. Il est naturel de s'attendre à une grande simplification du travail sur un système de traduction multilingue si l'on cherche à établir des équivalences sémantiques, plutôt que des correspondances, entre des langues particulières. Ce qui est important dans ces systèmes, c'est le contenu : le programme exclut les différences des formes grammaticales.

Certains chercheurs comptaient adopter les niveaux des structures profondes de Chomsky. Lamb voulait atteindre le niveau "sémémique" des grammaires stratificationnelles. Le CETA a opté pour l'analyse poussée au delà même des structures profondes de Chomsky. D'une part, les chercheurs s'intéressaient à la traduction multilingue. Ce groupe étudiait le russe, l'allemand et le japonais en tant que langues sources et cherchait donc à rendre les phases de transfert les plus simples possibles. D'autre part, la

traduction d'une phrase russe en français exige souvent la reprise de toute la construction syntaxique pour obtenir un résultat soigné. Comme le niveau syntaxique était jugé insuffisant pour effectuer un transfert maniable entre les langues de structure syntaxique très différente, l'analyse a été poussée jusqu'à l'interprétation de telles structures en termes de relations "prédicatives".[32] A ce niveau, la plupart des contraintes syntaxiques des langues étudiées disparaissent au profit de structures logiques. La transformation de chaque phrase source en une formule de langage pivot donnera lieu à des phrases cibles qui préservent le sens sans traiter les contraintes morphologiques et syntaxiques des langues naturelles.

En 1957 Yngve avait proposé l'organisation d'un système de TA en trois phases logiques: analyse monolingue, transfert bilingue, et génération monolingue.[33] Il s'ensuit que l'analyse doit fournir, pour toute unité de traduction (ici la phrase), un "descripteur structural" ne contenant aucune référence à la langue d'arrivée. Afin de faciliter le transfert, il faut donc rechercher un descripteur le plus universel possible. C'est enfin l'idée de base du "langage pivot". Les chercheurs du CETA étudiaient les systèmes

---

[32] Vauquois. La traduction, 8.

[33] Victor Yngve. "A Framework for Syntactic Translation." Mechanical Translation 4:3 (1957): 59-65.

linguistiques de Tesnière (les actants) ou de Mel'chuk (le modèle "sens-texte"). Finalement, ils choisirent un pivot "hybride", dans lequel les symboles grammaticaux et relationnels sont universels (temps abstrait, traits sémantiques, relations actancielles et circonstancielles comme cause, conséquence,..), mais où les symboles lexicaux sont empruntés à une langue naturelle.[34] Dans un système de TA fondé sur ce pivot hybride, le transfert se réduit à un transfert lexical, réalisé à l'aide d'un dictionnaire bilingue qui permet de déterminer le meilleur équivalent d'une famille de termes dans la structure produite par l'analyseur.

Ce langage intermédiaire serait idéal. Pourtant, selon Vauquois il a été nécessaire de limiter les perspectives d'une telle notation et de considérer un langage pivot accessible. Vauquois reconnaît l'impossibilité d'un langage pivot dans lequel on pourrait formuler tout texte de n'importe quelle langue.[35] L'analyse du CETA se borne donc à chercher le type de langage pivot qui sera un invariant pour un nombre limité de langues. Leur système est fondé sur le passage par ce "langage pivot" hybride. Il ne s'agit pas d'un langage

---

[34] Ch. Boitet. "TA et TAO à Grenoble... 32 ans déjà!" T.A.L. (revue semestrielle de l'ATALA). 33:1-2, Spécial Trentenaire (1992): 51.

[35] Bernard Vauquois. "Structures profondes et traduction automatique : Le système du C.E.T.A." Revue Roumaine de Linguistique. Tome XIII, No.2. (1968): 106-107.

universel. Puisque les chercheurs envisagent la réalisation de plusieurs programmes de traduction automatique dont la langue d'arrivée reste toujours la même, le langage pivot sera naturellement orienté vers cette langue privilégiée. D'ailleurs Vauquois écarte les textes "littéraires"; il limite le nombre de langues pour lesquelles le langage pivot est un invariant, et il circonscrit l'analyse par le contexte de chaque phrase. Si tout langage artificiel doit avoir son lexique et sa grammaire, le CETA ne donne pas de lexique propre au langage pivot. Le transfert se borne à substituer aux unités lexicales source, les unités lexicales correspondantes de la langue d'arrivée. Selon G. Veillon du CETA l'hypothèse de départ est dérivée des théories de Tesnières.[36] Ce dernier avait le désir de fonder une syntaxe générale, de concevoir une méthode d'analyse universellement applicable à toutes les langues. Veillon admet que le "pivot" du CETA est loin d'être de nature sémantique. La syntaxe est plutôt considérée comme une "structure syntaxique profonde commune à un groupe de langues naturelles".[37]

Vers 1961, Vauquois considère que pour les chercheurs à Grenoble le choix d'une conception de la traduction automatique se ramène à l'une des démarches suivantes :

---

[36] Lucien Tesnière. Éléments de syntaxe structurale. Paris : Klincksieck, 1969.

[37] Veillon. "Description", 8-17.

(i) Profiter des systèmes de première génération afin de réaliser un projet à court terme de traduction russe-français ;

(ii) Tâcher d'obtenir un système de T.A. donnant des résultats de meilleure qualité en rejoignant les équipes de deuxième génération ;

(iii) Se lancer dans des recherches originales afin de réaliser un système à un avenir lointain.[38]

Pour ne pas répéter les travaux de la première génération, et pour continuer sur un chemin réaliste, c'est dans la voie de la deuxième solution que se sont engagés les chercheurs du CETA.

3.3 PHASES DE TRADUCTION

Leur système est caractérisé surtout par l'emploi de plusieurs modèles successifs. A la différence des modèles de la langue qui représentent la langue elle-même, chacun de ces modèles représente un niveau de la langue. Il serait trop long d'entrer dans le détail et nous nous bornerons à indiquer la stratégie adoptée pour chacune de ces phases.[39] Puisque

---

[38] Vauquois. La traduction, 31-32.

[39] Le CETA poursuit la notion de modèle formel des niveaux de langue naturelle : "nous engageons dans la représentation de ces niveaux (de Lamb) au moyen de langages artificiels qui coincident avec les niveaux d'un système formel".

la phase d'analyse et la phase de synthèse présentent trop de difficultés pour être franchies en un seul pas, on les effectue par étapes successives. Chacune de ces phases est concrétisée par la représentation du texte dans un modèle qui correspond à un certain niveau de langue.

D'abord le texte est soumis à une préédition qui consiste à indiquer la séparation en phrases et à numéroter les "mots", à en calculer la longueur et à les cadrer dans un format standard. Le texte en langue de départ est préédité au moyen d'un programme qui réalise ainsi la décomposition en mots et y affecte des numéros d'occurrence consécutifs. Ces "mots", bien entendu, peuvent être des mots conventionnels ou bien des signes de ponctuation, des formules, des chiffres, etc.

La première phase se borne à une analyse morphologique du texte de départ. Le but de ce modèle est double : d'une part il faut décider si une segmentation est correcte ou non et d'autre part il faut calculer un certain nombre d'informations qui seront exploitées dans les modèles suivants. Les segmentations incohérentes sont rejetées et l'on substitue à la forme un ensemble d'informations en ce qui concerne son identification lexicale, ses traits grammaticaux, ses propriétés combinatoires sur le plan syntaxique et ses caractéristiques "sémantiques" ; ces informations constituent

---

Vauquois. La traduction, 33.

un "syntagme élémentaire".

Cette analyse s'effectue en deux temps. En premier lieu on recherche à partir de chaque mot du texte, appelé "forme", l'identification lexicale de cette forme ainsi que les diverses informations qui y sont attachées. Par exemple, supposons que l'on veuille obtenir les informations suivantes à partir du mot DISCUSSIONS :

- c'est un nom commun de genre féminin utilisé au pluriel
- ce nom est un nom d'action dérivé d'un verbe
- le verbe d'origine, qui sert d'identificateur lexicale est "DISCUTER"

Le procédé le plus rudimentaire, une méthode assez répandue à cette époque, consiste à classer toutes les formes et à établir ainsi un "dictionnaire de formes". La forme "DISCUSSIONS" se trouvera dans le dictionnaire au même titre que "DISCUTER", ainsi que toutes les formes conjuguées de discuter et les autres formes dérivées. Toutes les informations recherchées pour chaque forme apparaissent aussi dans le dictionnaire.

Dans le système du CETA, qui se sert des travaux de Lamb, la forme est considérée comme l'ensemble de deux composants : base et désinence.[40] Ce procédé permet la décomposition des

---

[40] S. Lamb et W. Jacobson "A High-speed Large Capacity Dictionary System." Mechanical Translation, vol 6, November 1961.

formes en un nombre quelconque de segments. En reprenant l'exemple précédent, l'analyse de "discussions" donne lieu au découpage : "discus + sion + s" où "discus" est dans le dictionnaire de bases et "sion" et "s" se trouvent dans le dictionnaire des suffixes et désinences.

La première phase de l'algorithme réalise la segmentation de la forme et la consultation du dictionnaire ; suivant la terminologie de Lamb, elle fait ainsi passer du niveau de la chaîne de "graphèmes" (le niveau du texte d'entrée sera appelé le niveau graphémique) au niveau de la chaîne de "morphes" (les graphèmes servent à représenter ces éléments de niveau immédiatement supérieur).[41] Pour accélérer le processus d'identification du morphe, Veillon a perfectionné DICADAP, se basant toujours sur les travaux de Lamb. Le programme transforme le dictionnaire initial pour le mettre sous forme arborescente. Veillon a construit le programme LEXICA qui vise à l'optimisation de cette forme. Le résultat de cette segmentation substitue à la forme d'entrée, les différents découpages de chaque forme au moyen des adresses de référence trouvées en regard de chaque morphe.[42]

---

[41] Sidney M. Lamb. Outline of Stratificational Grammar. Washington, D.C.: Georgetown UP, 1966.

[42] G. Veillon. Consultation d'un dictionnaire et analyse morphologique en Traduction Automatique. Thèse de 3ème cycle, Université de Grenoble I, juin 1962.

En deuxième lieu se réalise l'analyse morphologique proprement dite. Il s'agit d'employer les codes morphologiques pour rejeter les phrases illicites et pour déterminer les variables grammaticales de la forme lorsque la phrase est cohérente. Des catégories syntaxiques apparaissent comme résultats de l'analyse morphologique et sont utilisées au cours de l'analyse syntaxique. Le programme permet une exécution extrêmement efficace. Avec un dictionnaire de 12 000 bases, la segmentation s'exécute à la vitesse de 500 mots par seconde et l'analyse morphologique à la vitesse de 1 000 mots par seconde.

A l'entrée du modèle de l'analyse syntaxique est proposé une famille de suites de syntagmes élémentaires. Cette deuxième phase recherche les suites de syntagmes qui constituent une phrase. On élimine les suites de syntagmes qui ne constituent pas des phrases et fournit pour chaque phrase l'ensemble des structures syntaxiques compatibles avec la grammaire. Ainsi se précise la structure syntaxique de chaque phrase, structure dite "de surface", obtenue au moyen d'un modèle "hors-contexte".[43] L'unité de traitement est alors la phrase considérée comme chaîne de syntagmes

---

[43] Réalisée au moyen d'un analyseur syntaxique de type "Hors-contexte étendu" dont l'algorithme est une variante de l'algorithme de Cocke. Voir Bernard Vauquois: Syntaxe et interprétation Proceedings of the 1965 International Conference on Computational Linguistics, New York, mai 1965.

élémentaires ensuite transformée en structure arborescente.

Dans le système du CETA deux classes de mots sont distinguées : prédicats et descripteurs. Une phrase correspond à un graphe arborescent représentant la structure en langage pivot. Les mots sont mis en relations, symbolisées par des flèches ; chaque relation est caractérisée par une étiquette, associée au mot dépendant. Dans la phrase "Elle rend le livre à Pierre", "Elle" sera agent, "livre" objet, "Pierre" objet complémentaire (COMOBJ). A une telle formule correspond un graphe arborescent représentant la structure en langage pivot :

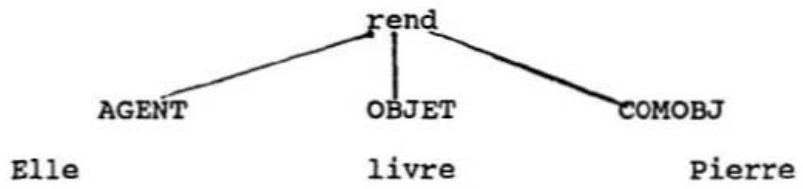


Fig. 1. Graphe arborescent pour la phrase "Elle rend le livre à Pierre".

L'analyse syntaxique associe donc à chaque phrase du texte une arborescence et correspond au niveau linguistique appelé "syntaxe de surface".

La caractérisation d'un langage constitue un problème majeur. Comment un langage qui peut contenir une infinité de

chaînes peut-il être représenté au moyen d'un processus fini? Pourtant, quant à l'analyse syntaxique, l'obstacle principal réside moins dans la description des phénomènes linguistiques que dans le rejet des formules parasites. Il faut décider pour toute chaîne à partir du vocabulaire terminal si cette chaîne appartient ou non au langage.

Dans l'analyse syntaxique en constituants immédiats, la phrase, unité syntaxique représentée par le symbole S, est décrit par le chaîne GN+GV; où GN et GV sont respectivement les symboles représentatifs des unités syntaxiques "groupe nominal" et "groupe verbal". Alors que dans l'analyse syntaxique sous forme de dépendances un ensemble de règles mettent en évidence les relations de gouverneur à dépendants. Par exemple, entre un article et un nom la relation de dépendance exprime que le nom est le gouverneur et l'article est le dépendant. Cette grammaire est fort incomplète puisque non seulement elle ne rend pas compte de toutes les structures de phrases, mais ce qui est plus important, elle accepte aussi une quantité considérable de phrases incorrectes au niveau des formes.

Le procédé utilisé par CETA considère la grammaire comme une donnée et produit des phrases en tant que résultat. Le problème qui intéresse les chercheurs voulant réaliser l'analyse syntaxique n'est donc pas celui de la caractérisation mais celui de la reconnaissance. Leur

stratégie d'analyse est celle de l'analyse ascendante. Il s'agit non seulement d'accepter ou de rejeter des phrases, mais de construire une arborescence à partir des symboles terminaux en construisant successivement les unités syntaxiques jusqu'au symbole S. Cette méthode permet de trouver toutes les solutions simultanément.[44]

La troisième phase sera une "interprétation" de cette structure syntaxique. En 1961, le contenu de cette interprétation, présenté sous forme de langage pivot, est imprécis. Il a fallu se limiter à substituer aux unités lexicales source, les unités lexicales correspondantes de la langue but. Dans certains cas, il sera nécessaire de modifier la structure. Le modèle a comme but de transformer chaque structure de "surface" en une ou plusieurs structures du langage pivot en utilisant une "grammaire transformationnelle" et de corriger éventuellement certaines sous-structures proposées par l'analyse syntaxique. A l'issue de l'analyse syntaxique, les phrases sources sont transformées en arborescences ; chaque sommet d'une arborescence comporte des informations (unité lexicale, valeurs de variables grammaticales, etc.) Une formule en langage pivot, formule associée à une phrase, se présente aussi sous forme d'une arborescence dont les sommets sont porteurs d'information. Le

---

[44] Vauquois. ANALECTES, 244-245.

processus de transition de la formule syntaxique de surface à la formule en langage pivot est donc une transformation d'arborescences.

Les structures syntaxiques, munies de leur code d'étiquetage, sont proposées au modèle de transfert qui calcule et choisit la formule intermédiaire correspondant à la phrase de départ. Des familles de mots sont alors substituées aux mots de langue de départ qui se trouvent dans le langage pivot, car celui-ci n'a pas de vocabulaire propre. Ce dernier est un langage artificiel dans lequel chaque formule a pour représentation une famille de phrases équivalentes de la langue de départ (par exemple: efficace, inefficace, efficacité, inefficacité, efficacement, inefficacement). A l'issue du modèle de transfert, une structure est attribuée à chaque phrase (dans le cas de phrases ambiguës : plusieurs structures). Cette structure comporte toutes les informations nécessaires au transfert à la structure de la langue d'arrivée. Il s'agit ici d'une transformation d'un arbre du type dépendance à un autre.[45]

---

[45] Comme nous l'avons indiqué, l'arbre de phrase-structure et celui de dépendance mettent l'accent sur deux aspects structuraux différents de la phrase : le premier sur les structures hiérarchiques des constituants et le deuxième sur des relations fonctionnelles entre le gouverneur et les dépendants. Un arbre de dépendance permet l'indication explicite des relations fonctionnelles avec lesquelles sont liés l'élément gouverneur et ses dépendants, ce qu'un arbre de phase structure ne peut pas faire, bien que celui-ci soit adéquat pour indiquer les degrés de connexité entre des constituants.

La génération d'une phrase française commence par l'élaboration d'une structure de surface. L'analyse et le transfert ont remplacé la phrase source par une formule en langage pivot à lexique français. L'arborescence qui représente cette formule est une arborescence étiquetée, où à chaque sommet est associée une étiquette complexe. Les phases successives sont consacrées respectivement à la génération de la structure syntaxique de surface des phrases et à la génération morphologique des mots. Le modèle de génération syntaxique reçoit les formules du langage pivot ; les codes syntaxiques des familles de mots équivalents français ont été ajoutés aux informations contenues en chaque sommet de la structure pivot. A partir des informations fournies par l'étiquette attachée à chaque sommet il faut déterminer de nouvelles variables liées à la syntaxe de la langue d'arrivée. On commence par calculer une fonction syntaxique puis une catégorie syntaxique pour chaque sommet. Ensuite, se succèdent les étapes de la synthèse : d'abord la synthèse de la construction syntaxique au moyen du modèle, dont la sortie est une suite structurée de syntagmes élémentaires de la langue d'arrivée. Enfin chaque syntagme est transformé en mots avec éventuellement préfixe, suffixe, désinence de la langue d'arrivée par le modèle.

Après le programme de construction morphologique, qui a été ici associée à la syntaxe, une post-édition effectue la

mise en ordre des mots conformément à leurs poids respectifs et réalise les diverses élisions et contractions, obtenant ainsi le texte français d'arrivée. On reprend la chaîne de mots de la langue d'arrivée dans le but d'exécuter les altérations morphologiques dues à la présence des mots voisins. En théorie la sortie sera donc le texte en langue d'arrivée. L'annexe à ce document présente un passage d'une telle traduction et fournit les résultats pour une phrase particulière de ce texte. (voir Annexes 1-5)

3.4 RÉSULTATS DU SYSTEME

Les chercheurs ont développé un système de grande importance axé sur le russe-français.[46] L'allemand-français et le japonais-français viennent s'y ajouter, et devront suivre les mêmes procédés de génération du français. Jusqu'aux tentatives japonaises dans les années quatre-vingt, ce système reste le seul exemple de réalisation de l'approche "pivot". Nous pouvons le rapprocher du système de l'Institut Textile de France, mais le "pivot" du système TITUS se limite à un langage très restreint.

Il est intéressant de considérer les résultats d'un tel système afin d'en souligner les faiblesses et les points

---

[46] Bernard Vauquois. "Présentation du Centre d'Études pour la Traduction Automatique (CETA) du Centre National de la Recherche Scientifique." T.A. Informations. 1966, 1.

forts. Les recherches du CETA avaient commencé en 1961. C'est en 1967 que nous voyons les toutes premières traductions russe-français obtenues par ordinateur. Quatre années plus tard on met la dernière main aux grammaires et aux dictionnaires. Les traductions russe-français obtenues et évaluées sur des volumes très importants pour l'époque étaient de qualité remarquable.[47]

En 1971, après l'expérience sur un échantillon tiré à des domaines variés (techniques aérospatiales, mathématiques, linguistique, physique nucléaire, thermodynamique), la traduction d'une masse de textes représentant plus de 400 000 occurrences de mots, le projet se termine. Un bilan effectué sur un échantillon de quatres textes d'origine différente (environ 15 000 mots) fait ressortir les sources de difficultés rencontrées au cours du processus : 42 pour 100 des phrases sont très compréhensibles. Les phrases compréhensibles ne sont pas des phrases équivalentes exactes puisque les modèles ne sont que des approximations. Cependant, elles sont acceptables. Il subsiste des erreurs dans le choix des articles et des ambiguïtés (des polysémies). Un petit nombre de ces phrases ne sont pas intelligibles à cause de leur longueur et le manque de renseignements concernant la

---

[47] Boitet. "TA et TAO à Grenoble...32 ans déjà!", 48.

structure superficielle en langue de départ.[48] D'après Vauquois "il est peut-être utile de garder, au niveau du langage intermédiaire quelques traces de la structure du texte d'entrée pour faciliter la génération." [49] De cette façon, même si l'on diminue l'ambition du langage pivot, il peut être possible de résoudre certains ambiguïtés. Par exemple les phrases : "Vous possédez ce stylo." ; "Ce stylo vous appartient." ; et "Ce stylo est à vous." donnent lieu à 3 formules différentes. Par contre : "Le secrétaire lit le journal." et "Le journal est lu par le secrétaire." sont reconnues équivalentes et sont représentées par une seule formule.

Reprenons chacun des modèles du système. Il faut aborder la question de l'utilité de la pré-édition. En général les textes soumis à la traduction automatique comprennent un vocabulaire et une grammaire déjà restreints. En outre, afin de résoudre les ambiguïtés qui s'élèvent, l'ordinateur peut poser des questions à son utilisateur. Néanmoins, l'effort requis par une telle pré-édition peut être compensé par le gain de temps dans la post-édition pour le texte d'arrivé. S'il s'agit de la traduction multilingue, le gain est encore plus important.

---

[48] Vauquois. La traduction, 142-148.

[49] Vauquois, La traduction, 155.

La séparation en niveaux des phases d'analyse et de génération permet de séparer les difficultés et de bien préciser les niveaux de langues en question. Dans la première phase sont effectués l'identification lexicale des mots source et le calcul de leur décomposition en morphes. Si un mot n'est pas reconnu, toutes les phrases qui contiennent ce mot sont rejetées. En conséquence, la première fois qu'un texte est soumis au processus, l'analyse morphologique est utilisée comme identificateur des mots inconnus. Au lieu de fournir des résultats possibles pour les mots, l'algorithme fournit une liste de tous ces mots et les mots correspondants sont incorporés dans le dictionnaire. L'incorporation des noms propres a tendance à encombrer inutilement le dictionnaire.

Enfin, ces travaux s'appuient déjà sur le principe de la séparation entre les données linguistiques et les programmes. Cela permet de clarifier et bien séparer les concepts qui interviennent dans la description des grammaires et dictionnaires de ceux qui interviennent dans l'algorithme. Au lieu de programmer directement pour chaque langue les opérations d'analyse, de transfert et de génération, les linguistes écrivent une fois pour toutes des programmes généraux qui constituent des modèles théoriques pour ces différentes phases. Ils construisent un ensemble de programmes auxquels le linguiste fournit les données linguistiques (grammaires et dictionnaires) ; ces données sont

rédigées dans un langage artificiel lié à cet ensemble de programmes.

Même si le modèle morphologique a son actif le bénéfice de l'efficacité, il reste néanmoins quelques aspects qui pourraient être traités de façon plus acceptable. L'unité de traitement est la forme. Les formes ambiguës donnent plusieurs solutions. C'est donc à l'analyse syntaxique de désambiguïser ces formules. Pourtant, cette analyse est longue et coûteuse. La plupart des ambiguïtés pourraient être éliminées plus simplement par l'analyse d'un contexte restreint.

L'analyse syntaxique se réalise d'une manière semblable. Une phrase qui ne peut être analysée complètement est en panne et ne fournit aucun résultat. Ceci a pour effet de perdre du temps puisque certains aspects de l'analyse sont remis en cause au cours des phases successifs. De plus, le modèle d'analyse syntaxique en constituants fournit un trop grand choix de solutions. Ce défaut vient du fait que l'on ne connaît pas la quantité d'informations qu'il faut attribuer aux unités lexicales pour que le modèle ait un pouvoir de sélection suffisant. L'unité d'analyse se limite à la phrase du texte de départ. Cette limitation interdit par exemple de trouver les antécédents ; elle oblige à accepter le risque de traduire avec des contresens des formules pour lesquelles un contexte moins restreint aurait été nécessaire.

Les algorithmes de consultation de dictionnaires, les analyses morphologiques et les analyses syntaxiques ont tous fait l'objet de recherche en efficacité maximale. Pourtant, la vitesse de traduction sur ordinateur IBM 7044 reste entre 4 000 et 5 000 mots à l'heure suivant les textes.[50] C'est trop lent pour qu'un tel système soit économiquement viable sur une grande échelle de production.

3.5 ORIENTATIONS SOUHAITABLES

Vauquois indique plusieurs orientations souhaitables sur le plan linguistique :

(i) Spécification des traits les plus pertinents et les plus efficaces.

(ii) Détermination des niveaux de langue les plus pratiques pour le transfert et l'élaboration d'un langage pivot.

(iii) Organisation des grammaires pour éviter le traitement des phrases simples par un processus trop complexe.

---

[50] Vauquois. La traduction, 142.

D'après Vauquois :

> ...on essaiera des grammaires qui pourront tenir compte d'une telle pré-édition sans que celle-ci soit nécessairement présente ; la différence se verra à la qualité du résultat et au temps de calcul passé pour l'obtenir. Engagé dans cette voie, le Groupe d'Étude pour la Traduction Automatique a commencé par réaliser les outils informatiques nécessaires à une telle orientation.[51]

Boitet (CETA) envisage un développement de la recherche vers l'accès à des niveaux de langue plus proche de la "compréhension".[52] La puissance sémantique du langage pivot est caractérisée par la capacité qu'il a de reconnaître les phrases équivalentes lorsque celles-ci diffèrent seulement par leur construction syntaxique. L'étiquetage est déjà l'ébauche d'une sémantique. Cette opération se propose d'attribuer une signification aux relations grammaticales des différents membres de la phrase. Alors se dessine déjà le lien entre les travaux effectués en vue de la TA et l'intelligence artificielle.[53]

---

[51] Vauquois, La traduction, 147-148.

[52] Voir Boitet : "...espérons que l'utilisation de méthodes plus heuristiques et d'informations plus sémantiques (gnosto-encyclopédiques), en combinaison avec des approches plus classiques, mais, me semble-t-il, nécessaires, permettra d'améliorer de façon substantielle la qualité des traductions produites par les systèmes de TA."
Ch. Boitet. "Méthodes sémantiques en Traduction Automatique." T.A. Informations. 1 (1976): 39.

[53] D'après Boitet, le système du CETA, comme celui de l'Université de Georgetown, n'utilise qu'une sémantique élémentaire. Cependant, à son avis, muni du modèle d'éti-

Reconnaissant qu'aucune théorie linguistique ne répond à tous leurs problèmes, les chercheurs du CETA ont dû modifier leurs systèmes à des fins pratiques. Ils ouvrent une voie toute nouvelle, profondément inspirée de la théorie linguistique et de la théorie des langages formels modernes. Dès lors ils insistent sur la création des "langages spécialisés", aussi appelés LSPL (langages spécialisés pour la programmation linguistique).[54] Toutefois, les efforts fournis par les chercheurs du CETA n'ont pas toujours été couronnés de succès. Le rapport ALPAC laissait planer un scepticisme sur la possibilité de réalisation de la TA. Vauquois et son équipe ont démontré la fausseté des conjectures pessimistes initiales des années soixante. Même si l'idéal inaccessible d'une traduction automatique de qualité parfaite avait dû être remplacé par des objectifs plus réalistes, les résultats furent impressionnants.[55]

---

quetage qui transforme le structure en une structure pivot, c'est un système de deuxième, presque de troisième génération.

[54] Les LSPL facilitent le travail des linguistes et des lexicographes de façon qu'ils ne leur faut pas s'inquiéter de certains problèmes informatiques.

[55] "Dès la fin des années soixante, les traductions russe-français obtenues et évaluées sur des volumes très importants pour l'époque (plus de 400,000 mots, soit 1600 pages standard) étaient de qualité impressionnante, à tel point que, tout nouveau venu au CETA en octobre 1970, je me demande franchement s'il ne valait pas mieux chercher un autre domaine de recherche, tout semblant déjà avoir été trouvé!" Vauquois. ANALECTES, 1.

En 1971, un changement du matériel et ainsi du logiciel a été suivi de l'abandon du système de CETA. Ce système est abandonné un faveur d'un nouveau plan - celui du Groupe d'Étude pour la Traduction Automatique (GETA). Dans la deuxième partie de notre étude, nous examinerons la nouvelle orientation des activités du GETA depuis 1972. Les chercheurs s'en tiennent à concevoir et à réaliser des modèles d'analyse, de traduction ou de génération de langues naturelles qui serviront d'outils pratiques pour la traduction automatique assistée par ordinateur.

4.0 LE GETA ET LA TA DU RÉVISEUR

4.1 CRÉATION DU GETA

A partir de 1971 le CETA, désormais désigné "GETA" (Groupe d'Étude pour la Traduction Automatique), n'est plus reconnu comme laboratoire propre du CNRS. Aux effets du rapport ALPAC, publié par l'Automatic Language Processing Advisory Committee en décembre 1966, viennent s'ajouter les "événements" de mai 1968, aboutissant à une perte de crédibilité du domaine et à la division de l'équipe du CETA en trois nouvelles unités. L'une, gérée par J. Rouault, s'attaque aux problèmes de linguistique théorique et d'informatique documentaire. La seconde, dirigée par G. Veillon, travaille d'abord sur des outils d'analyse interactive du français, puis évolue vers l'intelligence artificielle, et enfin vers la robotique. Seul le GETA, sous la direction de B. Vauquois, poursuit le problème global de la TA. Le ministre de la Défense, envisageant la traduction des documents russes, décide de soutenir les efforts du GETA. Le CNRS projette aussi de relancer une recherche fondamentale, dans le cadre du GETA, laboratoire "associé" au CNRS. L'informatique évoluait, les systèmes interactifs apparaissaient. L'ordinateur IBM 7044 est remplacé par un IBM 360/67, la première machine à mémoire virtuelle. Les nouveaux concepts semblent promettre une réalisation convaincante sur

le plan scientifique.

4.2 TRADUCTION DU RÉVISEUR

La réflexion sur des expériences menées au CETA dans les années soixante a amené le GETA à projeter des logiciels pour satisfaire aux exigences des langues naturelles et répondre aux critiques des systèmes du CETA. De vaines tentatives en vue de la traduction automatique de haute qualité ont amoindri le prestige de la "TA" qui cède le pas à la "TAO" (Traduction Assistée/Automatisée par Ordinateur) pour indiquer des possibilités élargies d'informatisation de la traduction. Dès lors la traduction visée par les systèmes du CETA s'appelle la "TA du veilleur", destinée à produire des traductions brutes, sans révision, pour relever le "sens" de n'importe quel texte.[56] Plus tard il s'agira de la "TA du réviseur", de la "TA interactive", ou la "TA du rédacteur", et enfin de la

---

[56] D'après Vauquois la "TASR" (Traduction Automatique Sans Révision) "De 1960 à 1970, le CETA, dirigé par le Pr.B. Vauquois, a travaillé sur la TAO du veilleur, arrivant à construire un système russe-français impressionnant, de qualité suffisante pour permettre à un spécialiste de comprendre des textes russes scientifiques et techniques de domaines variés (physique nucléaire, chimie, linguistique, sciences de l'espace)." Voir Ch. Boitet. "La TAO à Grenoble en 1990 : Présentation générale." La TAO à Grenoble en 1990. Grenoble: IMAG, 1990, 1.

"TA(O) du traducteur".[57]

De 1971 à 1981, le GETA, s'orientant vers la "TA du réviseur", système qui vise à produire automatiquement des traductions brutes destinées à être peaufinées, expérimente une nouvelle méthodologie. Selon Boitet, le but est de produire automatiquement des 'premiers jets' dont la révision rapide par des professionnels fournirait la traduction à un rythme accéléré et à moindre coût.[58] Le GETA se porte sur l'approche transfert multiniveau et réalise un générateur de systèmes de TAO (Ariane-78) permettant d'écrire des maquettes de TAO multilingue au moyen des "LSPL". Il a fallu avoir recours à des grammaires dynamiques modulaires et à des techniques de programmation heuristique pour réaliser ce premier générateur de systèmes de TAO.[59]

---

[57] Ou THAO (Traduction Humaine Assistée par Ordinateur), dans laquelle on fournit un "poste de travail" adapté au traducteur-réviseur.

[58] Vauquois l'a appelé la "Traduction Automatique Avec Révision" (TAAR). Voir Boitet. "La TAO à Grenoble en 1990", 1.

[59] Le GETA s'est également intéressé à la TAO du "traducteur", c'est-à-dire aux outils bureautiques isolés ou aux "postes de travail de traducteur", bien adaptés aux petits volumes de textes hétérogènes, qu'il est globalement exclu de traduire automatiquement. Ces postes de travail fourniraient aussi la révision de traductions brutes. Boitet justifie la position du GETA envers le TAO du traducteur : "il s'agit d'un complément indispensable aux outils de traduction automatique, et de nombreux industriels y travaillent. C'est d'ailleurs pourquoi cet aspect a été relativement peu développé au GETA." Boitet. "La TAO à Grenoble en 1990", 17.

Cette problématique évoluée a exercé un effet considérable sur la méthode. Ayant observé la complexité de grosses grammaires avec leurs milliers de règles, Vauquois se met à la recherche d'une organisation plus modulaire. Enfin, ayant vu les problèmes d'une approche combinatoire, il se fixe sur une programmation linguistique heuristique.[60]

Une étude détaillée des systèmes étrangers a été conduite. Le GETA les divise en deux classes :

(i) systèmes dérivés de List Processing Language (LISP), Réseaux de Woods,[61] Programmar de Winograd,[62] dont

---

[60] Dans chaque type de modèle, des méthodes combinatoires ou heuristiques sont utilisées. Boitet précise: "une méthode est "combinatoire" si on commence par énumérer toutes les solutions avant de choisir, et heuristique si la poursuite ou l'arrêt de la construction d'une solution est guidé par une estimation." Ch. Boitet. Méthodes sémantiques, 39.

Le langage ATEF de Chauché (voir 62-64) comporte un "graphe de contrôle" aux noeuds duquel sont attachées des grammaires transformationnelles. Le non-déterminisme de l'ensemble de règles ajouté à la possibilité de récursion sur des sous-arbres permettent l'adoption des stratégies "heuristiques". Le "moteur" d'ATEF traite successivement chaque occurrence du texte en examinant toutes les analyses possibles, sauf si des fonctions spéciales sont appelées. Des fonctions de contrôle permettent au linguiste de mettre en oeuvre des heuristiques. Il peut contrôler la rétrogression (backtrack), produire un résultat partiel ou décider qu'une borne de phrase est atteinte. Si le linguiste n'utilise pas ces actions, l'analyseur écrit est combinatoire.

[61] William Woods. Transition Network Grammars for Natural Language Analysis. CACM 13/10, oct.1970, 591-606. CACM (Communication of the Association for Computing Machinery)

les grammaires sont rédigées par les utilisateurs. Des langages de programmation sont surtout adaptés à l'analyse syntaxique.

(ii) systèmes à "algorithme incorporé" (systèmes-Q de Colmerauer), dans lesquels les grammaires rédigées par l'utilisateur sont des "données"; ces systèmes comprennent également la compilation des dictionnaires.

Quant aux systèmes du premier type, ils laissent à l'utilisateur la possibilité de choisir son algorithme. Cependant, l'on éprouvera de la difficulté à séparer la partie "statique" de la grammaire de l'enchaînement des règles. C'est donc vers des "systèmes-Q" que s'orientent les chercheurs. Ces derniers apportent une réponse à bien des critiques des systèmes de deuxième génération. Il s'agit de transducteurs d'arborescences (et non plus d'accepteurs) complètement intégrés. C'est-à-dire que les données lexicales et grammaticales sont incorporées dans le système. La décomposition d'une grammaire volumineuse, comme celle du CETA, en une séquence de grammaires réduites plus maniables

---

[62] T. Winograd. "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language". AI-TR-17, MIT, Cambridge, Mass., jan. 71.
AI-TR (Artificial Intelligence Technical Report)
"Procedural Model of Language Understanding". Computer Models of Thought and Language. Shank & Colby, eds, Freeman, San Francisco, 1973, 152-186.

constitue un grand progrès dans le cadre d'une réalisation pratique du système. Bien que les systèmes-Q soient séduisants par la commodité de leur mise en oeuvre, leurs grammaires comportent un nombre de règles encore trop grand pour être pratiques. De plus, ils imposent la rigidité de l'algorithme choisi et excluent l'introduction de procédés heuristiques pour guider et éventuellement modifier cet algorithme.[63]

4.3 LANGAGES SPÉCIALISÉS ATEF, CETA, SYGMOR ET TRANSF

Le principal choix effectué dans les systèmes du GETA réside dans l'utilisation formelle d'arborescences étiquetées. La facilité de choisir entre les étiquettes, ainsi qu'entre les structures possibles, donne à ces systèmes de vastes champs d'applications. C'est sur ces notions que penche J. Chauché lorsqu'il rédige sa thèse doctorale.[64] Ses études au sein du GETA portent sur la définition et sur l'implémentation

---

[63] Voir A. Colmerauer. "Les systèmes-Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur". Projet TAUM, Université de Montréal, janvier 1971.
Les systèmes-Q ne permettent qu'un enchaînement séquentiel des grammaires (pour traiter chaque phénomène linguistique). Le système CETA permet un enchaînement arborescent (en hiérarchie), où le choix de la grammaire à appliquer s'effectue selon la vérification de conditions.

[64] J. Chauché. Transducteurs et arborescences. Étude et réalisation de systèmes appliqués aux grammaires transformationnelles. Thèse d'État, Grenoble, 1974.

du système ATEF (Analyse et Transduction en États Finis) et du système "CETA" (Contrôle Et Transduction d'Arborescences). Il envisage l'utilisation d'ATEF dans la formation des analyseurs morphologiques et de CETA pour les analyseurs structuraux et toutes les phases transformationnelles.

A Grenoble, le système ATEF est consacré à l'analyse morphologique doublée d'une première étape dans l'analyse syntaxique d'un certain nombre de langues naturelles : russe, français, allemand et japonais. Ce premier logiciel est un transducteur de chaîne à arborescence. Le système CETA est un deuxième logiciel, transducteur d'arborescence à arborescence. Un troisième logiciel permettrait de construire la chaîne des mots du texte de sortie à partir d'une arborescence (la génération du système SYGMOR). Lors du transfert de structures profondes il importe d'employer un dictionnaire bilingue. Un quatrième logiciel, TRANSF, s'applique à cette substitution lexicale. Chaque logiciel offre un métalangage d'écriture pour les grammaires et les dictionnaires. Ces systèmes impliquent l'écriture de grammaires par les utilisateurs spécialistes qui auront la capacité de guider la stratégie au moyen de fonction de contrôle (ATEF) ou d'un graphe de contrôle (ROBRA).

Une arborescence étiquetée est un ensemble de points munis d'une structure : d'une relation possédant des propriétés. Chacun de ses points est muni d'une étiquette.

L'étiquette est composée d'un ensemble de renseignements associés à ces points. La décoration est exprimée au moyen de variables qui se répartissent selon les niveaux d'interprétation :

(i) Classes morpho-syntaxiques - le niveau le plus superficiel attribue à chaque sommet de l'arborescence une classe qui représente les propriétés combinatoires sur le chaîne des mots (verbe, nom, ponctuation).

(ii) Fonctions syntaxiques - représentent le lien entre le subordonnant et le subordonné obtenu au moyen de critères formels, tels qu'accords en genre, nombre, personne (sujet, 1er objet, attribut du sujet / de l'objet direct).

(iii) Relation logico-sémantique - une information de type universel sur le nature de dépendance existant entre les mots ou les groupes de mots (but, possession, cause, résultat, instrument, moyen).

Ainsi, la figure suivante représente une arborescence étiquetée :

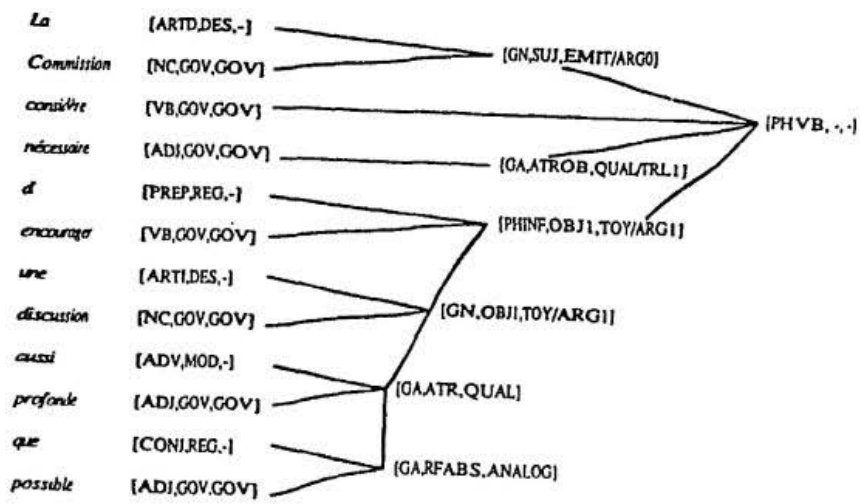


Fig. 2. L'arborescence étiquetée de la phrase "La Commission considère nécessaire d'encourager une discussion aussi profonde que possible.", dans laquelle on a noté les variables associées aux niveaux d'interprétation.

Le système ATEF a pour but la transformation d'une chaîne de mots d'entrée en une arborescence étiquetée, chaque mot de la chaîne d'entrée pouvant conduire à un ou plusieurs points de l'arborescence finale. Pour obtenir une arborescence étiquetée à partir d'une chaîne de mots à l'entrée (arborescence qui peut être manipulée par le système CETA), le système ATEF utilise un "dictionnaire" et une grammaire d'états finis. Le dictionnaire est un ensemble de segments (suite de caractères) auxquels sont associées une étiquette, une référence de traitement, et une référence d'unité lexicale. La référence de traitement spécifie le traitement particulier qui devra être associé à ce segment. Le système ATEF possède un maximum de six dictionnaires sans compter celui qui sert à indexer les tournures idiomatiques figées et invariables.

Le texte d'entrée se présente comme une chaîne de caractères ; le blanc sépare deux "formes" consécutives. Le "moteur" d'ATEF traite successivement chaque occurrence en examinant toutes les analyses possibles, à moins que des fonctions spéciales soient requises. L'analyse des formes est dans l'ordre de lecture de gauche à droite.[65] Chaque forme

---

[65] Le découpage de chaque forme à partir de la gauche ou de la droite est un paramètre laissé au choix de l'utilisateur.

est analysée par consultation des dictionnaires et application des règles de grammaire. Le résultat de l'analyse est donné par une arborescence dont chaque sommet est étiqueté par un "masque de variables". Le système ATEF possède des valeurs d'unité lexicale pour décorer les étiquettes du sommet associé au texte (ULTXT), du sommet associé à chaque phrase (ULFRA), à chaque occurrence (ULOCC) et à chaque mot composé (ULMCP). Ainsi, pour la phrase suivante : "Cette porte ferme mal.", on obtient l'aborescence de la figure 3, où la partie __ du masque contient les valeurs des variables calculées d'après la grammaire et les dictionnaires établis par le linguiste.

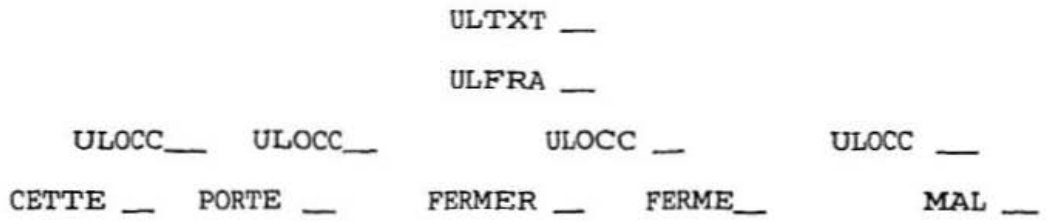


Fig. 3. L'arborescence pour la phrase "Cette porte ferme mal."

Chaque forme initiale peut fournir plusieurs découpages pour aboutir aux mêmes résultats. Seule une étude linguistique de ces phénomènes permet de prendre une décision sur la stratégie à adopter. Aussi, pour la forme "FERME"

obtient-on le verbe "FERMER" et le nom/adjectif/adverbe "FERME". Le système peut analyser, par exemple, tous les découpages possibles à partir d'un segment donné en commençant par les découpages comprenant le plus de caractères. Le résultat final de cette méthode est une arborescence étiquetée. La solution d'une phrase consiste en une suite d'étiquettes (une pour chaque mot de la phrase), chacune représentant une interprétation d'un mot de la phrase. La phrase n'est pas encore structurée, seules les ambiguités sont séparées.

Dans le système CETA une grammaire n'est pas l'ensemble de toutes les règles. Le linguiste élabore la stratégie d'analyse ou de génération. Il construit des sous-grammaires; chaque sous-grammaire est une liste ordonnée de règles. Ces règles s'appliquent simultanément sur toutes les régions de l'arborescence où les conditions d'application sont satisfaites. Le système organise lui-même le retour arrière (toujours minimum) chaque fois qu'une suite de sous-grammaires se trouvent dans une impasse.

Le langage spécialisé CETA, comme nous l'avons indiqué, est un transducteur arborescence-arborescence. Ce système permet d'écrire et de simuler une grammaire transformationnelle. Il manie des arborescences étiquetées produites par le système ATEF. Pour créer une grammaire transformationnelle à l'aide de cet ensemble, deux éléments

complémentaires sont nécessaires : l'ensemble de règles et l'ensemble des grammaires et la définition de leur enchaînement. Le système exploite des règles de grammaire et la gestion des transformations effectuées sur l'arborescence d'entrée. L'algorithme peut fonctionner selon des modes différents laissés au choix de l'utilisateur.

Le langage spécialisé CETA adopte le point de vue de l'utilisateur. C'est un langage d'écriture de systèmes transformationnels. Des grammaires transformationnelles sont attachées aux noeuds d'un graphe de "contrôle". Avec les nouveaux logiciels, une séparation des modèles par niveaux, comparable aux systèmes du Centre D'Études pour la Traduction Automatique, n'est plus nécessaire. L'objectif de modularité est atteint, car si l'ensemble de règles peut être énorme, chaque grammaire réduite ne comporte qu'un petit nombre de règles. Pourtant, pendant la première moitié des années soixante-dix, il ne s'agit que d'une maquette qui sert de banc d'essai à tous les nouveaux outils. Le développement de cette maquette et la mise en pratique des idées exposées contribuera à l'évolution de ces outils. La deuxième partie des années soixante-dix sera marqué par la réalisation du premier "générateur de systèmes de TAO" : ARIANE.

Le langage spécialisé TRANSF effectue le transfert lexical entre les mots de la langue de départ et ceux de la langue d'arrivée. Les données linguistiques sont livrées par

un dictionnaire bilingue. L'algorithme de ce système examine tous les sommets de l'arborescence qui lui est présentée. C'est la première étape de la phase de transfert.

Le langage spécialisé SYGMOR s'emploie pour faire passer de la structure de syntaxe superficielle au texte définitif en langue d'arrivée. C'est un transducteur arbre-chaîne, servant à "aplatir" l'arborescence d'entrée. Les sommets de l'arborescence qui vont figurer dans la chaîne sont sélectionnés. De cette chaîne de masques de variables se construit une chaîne de caractères représentant le texte final. Les données linguistiques qu'il faut fournir à SYGMOR pour réaliser un modèle de génération morphologique consistent en une grammaire et des dictionnaires monolingues. Ainsi, pour la variable "Unité Lexicale" PRODUIRE, le dictionnaire fournit la chaîne PRODUI pour la base du verbe conjugué, la chaîne PRODUCT pour la base des mots dérivés comme PRODUCTION, PRODUCTEUR, etc. Le linguiste a la liberté de choisir et de spécialiser les divers dictionnaires (dictionnaires pour unités lexicales, pour préfixes, désinences, suffixes, etc.).

4.4 SYSTEME ARIANE-78

Dès que ces nouveaux moyens linguistiques seront disponibles, Vauquois se mettra à améliorer l'organisation linguistique des traitements. Dans la séparation des "niveaux

d'interprétation" linguistiques avec les systèmes précédents il a fallu faire face à des problèmes importants. Par contre, des analyseurs faisant appel à tous les niveaux en même temps permettraient des retours en arrière ou garderaient toutes les possibilités jusqu'à la fin de l'analyse. Vauquois a conçu une technique inspirée du "tableau noir" pour la TAO, exposée pour la première fois à une réunion du groupe Leibniz.[66] L'approche transfert multiniveau, une nouvelle méthodologie de programmation linguistique est proposée.[67] Le GETA développe un logiciel permettant d'écrire des systèmes de traduction automatique, de produire des traductions et de les réviser sur l'écran. En 1977 le GETA participe au colloque "Franchir la barrière linguistique" à Luxembourg. Ses chercheurs y

---

[66] La Traduction Automatique était si souvent critiquée à cette période qu'il est devenu essentiel de réaliser des systèmes exploitables, d'une portée limitée, dans un proche avenir. De plus, il importait de poursuivre des recherches vers des objectifs plus ambitieux pour un avenir plus lointain. Ce travail dépasse la capacité d'un seul laboratoire. En 1975 c'est le groupe LEIBNIZ, un ensemble de 8 laboratoires européens, qui se chargera de ce projet.

[67] Trois niveaux ont été définis :
(i) celui des classes syntaxiques - le plus proche de la structure de surface
(ii) celui des fonctions syntaxiques - une relation entre deux sommets (subordonnant et subordonné)
(iii) celui des relations logico-sémantiques - censées constituer une information de type universel sur la nature du rapport de dépendance entre les mots ou les groupes de mots.
J.-Ph. Guilbaud. Descripteur Linguistique multiniveau et génération de texte en Ariane-78. Présenté pour la Journée ATHENA sur la Traduction Automatique. Université de Liège, 5 mai, 1987, 3.

exposent leur nouveau système de TAO qui, lors de l'intégration de toutes ses parties, sera baptisé Ariane-78, puisque la première version est disponible en 1978. L'environnement informatique de cette nouvelle approche est expérimentée sur divers couples de langues, dont principalement le russe-français. Grâce à des langues spécialisées et à un environnement interactif, l'informatique mise au service de non-informaticiens, permet de travailler de façon autonome.

Il s'agit d'un projet multilingue, une même analyse pouvant être branchée à plusieurs transferts, et une même génération à plusieurs transferts. Par conséquent, l'analyse doit être indépendante de la langue d'arrivée et, inversement, la génération ne peut pas dépendre d'une langue de départ particulière. Seule la phase de transfert est commune à un couple de langues donné. Le GETA aurait pu se servir des structures "pivot" comme auparavant. Néanmoins, l'expérience précédente avait évoqué deux problèmes importants : les traductions fournies par l'approche pivote ne sont que des périphrases, d'ailleurs, les informations de surface ayant été effacées, il est impossible de communiquer le parallélisme stylistique cherché. Reste également le problème de "tout ou rien": si l'analyse n'arrive pas à fournir un résultat complet au niveau sémantique, seule une traduction mot-à-mot est possible. Ainsi, selon l'approche précédente, ce qui n'était pas prévu n'était pas analysé, mais tout simplement traduit

mot-à-mot. Donc Vauquois donne à la TAO le but supplémentaire de chercher à tout traduire. Le système ROBRA, un ensemble de grammaires et un procédé d'enchaînement conditionnel de ces grammaires, fournit un résultat dans tous les cas. L'arborescence est donné comme "sous-produit" reflétant les dérivations dans la grammaire. L'arborescence d'entrée est traitée par des grammaires qui fournissent des arborescences résultats. La "transduction" ne se termine que lorsque l'arborescence d'entrée a été transformée par une séquence de grammaires dont la dernière grammaire est en position de feuille dans l'hiérarchie.

Ce nouveau système de programmation repose sur les quatre LSPL (voir p.62-64). Ils permettent d'écrire les dictionnaires et les grammaires qui constituent les différentes étapes du processus de traduction automatique. Le système repose également sur un moniteur interactif, ainsi qu'un système de traitement des textes et des corpus.[68] Le moniteur conversationnel a pour but de fournir au linguiste un outil complet et d'utilisation relativement simple, pour la mise au point d'une application de TAO, ainsi que pour son exploitation. Le moniteur prend l'utilisateur complètement en charge, du début à la fin de son travail avec l'ordinateur.

---

[68] Ch. Boitet, Pierre Guillaume, et Maurice Quézel-Ambrunaz. "Ariane-78, an integrated environment for automated translation and human revision". Proc. COLING-82, North-Holland, Ling. series 47, Prague, 1982, 19-27.

On ne doit connaître que les LSPL, l'architecture du système et l'éditeur. En plus d'un algorithme qui effectue le traitement proprement dit, il comporte un ensemble de programmes de visualisation qui permet à l'utilisateur de suivre le fonctionnement de l'algorithme à divers degrés de précision et ainsi de repérer facilement les erreurs dans sa grammaire et ses dictionnaires. L'interface réalisée par un moniteur interactif offre plusieurs modes d'utilisation :

(i) exécution en pas-à-pas.

(ii) production de traductions pour une liste arbitraire de textes.

(iii) révision manuelle, sur l'écran multifenêtre, avec quelques fonctions spécialisées, et l'accès éventuel à un dictionnaire modifiable par le réviseur.

Une fois entré sous le moniteur conversationnel, qui fonctionne comme une hiérarchie de menus correspondant à autant de sous-environnements, l'utilisateur répond à des questions. Pour chaque commande, il peut obtenir des explications détaillées sur la liste de paramètres qui doivent l'accompagner. A tout moment, il est possible de remonter d'un niveau ou sortir d'Ariane et obtenir la trace de la session.

Un langage de commande permet à l'utilisateur de mettre

à jour chaque partie du modèle ou bien d'exécuter tout ou partie d'un processus de traduction sur le texte de son choix. On recherche "un contrôle de ces outils par d'autres systèmes pour interrompre, lancer ou modifier leur fonctionnement à tout moment opportun." [69] Le GETA a choisi d'étudier des systèmes permettant une traduction brute de l'intégralité du texte soumis, sans aucune intervention manuelle. Malgré l'éventualité, ou bien l'objectif de l'automatisation partielle, ce système est toujours en mesure de fournir des traductions "brutes." [70] Vauquois a estimé que même si les syntagmes et les fonctions syntactiques, ainsi que les relations logiques, se conservent entre langues d'une même famille, il faut, pour accéder aux invariants entre groupes, aller jusqu'aux relations sémantiques.[71] Le GETA s'intéresse

---

[69] Vauquois. La traduction, 169.

[70] Boitet explique : "Machine (Assisted) Translation systems, or M(A)T systems for short, are a subset of the Computer-Aided Translation (CAT) systems. Their aim is not to assist a human translator while he is translating, but rather to produce either a "raw translation", input to one or more revision cycles, or a "crude translation", which is not supposed to be revised, but to be informative." Ch. Boitet. "Software and Lingware Engineering in Modern M(A)T Systems." Computational Linguistics : An International Handbook on Computer-Oriented Language Research and Applications. Berlin: Walter de Gruyter, 1989, 670.

[71] Ch. Boitet. Le logiciel ARIANE 78.5 du G.E.T.A. : Principes généraux, applications actuelles et futures. 5e Congrès national sur l'information et la documentation. 8-10 juin 1983. Grenoble: Information, Documentation, Transfert des Connaissances, 1983, 124.

au multilinguisme. On cherche donc à analyser un texte jusqu'au niveau profonde.

Les structures "multiniveaux" d'Ariane sont aussi un moyen d'implémenter des stratégies "robustes". Les systèmes de TAO de première génération résolvaient les ambiguïtés, autant que possible, lors de leur apparition. Aucune possibilité de retourner en arrière. Cependant, la deuxième génération de systèmes de traduction automatique aborde les ambiguïtés de façon plus avancée: on peut "traiter" les ambiguïtés en les représentant au moyen de la structure manipulée par le système de règles. C'est effectivement la technique employée dans les systèmes du GETA. Chaque phrase et chaque mot possèdent une structure linguistique. Le système Ariane peut analyser les phrases automatiquement en leur associant un descripteur de structure qui rend compte de plusieurs niveaux d'inter-prétation linguistique. A l'encontre des systèmes d'analyse par niveaux successifs, il est possible de progresser simultanément dans tous les niveaux et d'utiliser les résultats partiels obtenus dans un seul niveau pour résoudre les problèmes rencontrés dans les autres. Lors d'une analyse incomplète d'un énoncé au niveau d'interprétation le plus élevé,[72] on peut traduire le morceau

---

[72] "Le niveau profond...demeure inchangé de la langue source à la langue cible..." Guilbaud. Descripteur Linguistique, 3.

fautif au niveau inférieur. Par exemple, moyennant ce procédé, on peut se servir de l'information sémantique pour désambiguïser au niveau syntaxique :

Jean poussa Marie par la fenêtre.

Ses remords furent intenses.
Ses funérailles furent pénibles.

Le recours à ces structures multiniveaux, sous la forme d'un arbre décoré est nécessaire pour représenter les ambiguïtés non-résolues, car il permet d'éviter l'explosion combinatoire sur des énoncés longs. Pour lever l'ambiguïté qui subsiste dans les phrases ci-dessus, il faut rechercher l'antécédent dans la phrase précédente. Cela pourrait être obtenu sans difficulté en rajoutant une règle d'analyse. En éloignant le problème du "tout ou rien", le GETA fait un pas important vers un processus de traduction dans sa totalité, évitant le mot-à-mot. Selon Vauquois : "c'est jusqu'à présent la seule technique de traduire non pas phrase par phrase, mais paragraphe par paragraphe, voire page par page." [73] Cependant le désavantage qui en résulte, c'est qu'il est impossible d'écrire les règles dans l'ignorance des ambiguïtés éventuelles; la programmation devient plus fragile. De plus, il se trouve qu'une règle ne soit pas valide dans tous les cas du langage courant et son application aboutira parfois à des

---

[73] Vauquois. ANALECTES, 7.

erreurs.

Tout ce système existe en deux versions, anglaise et française. Il est implanté en France et à l'étranger, ayant servi à de nombreuses expériences, de type recherche (maquettes) ou développement (applications de grande ampleur). Dans les années quatre-vingts, le GETA a poursuivi le développement d'une application russe-français. Pourtant, l'application la plus développée est un système russe-français conçu pour la traduction de certains résumés du VINITI (le "Referativnyij Zhurnal") portant sur les sciences spatiales, la chimie et la métallurgie. D'autres applications sont plutôt d'étude : français-anglais (télécommunications), anglais-malais (chimie), analyses de l'allemand et du portugais. De nombreuses maquettes ont été également développées, à l'occasion de thèses de stagiaires, qui ont permis de valider certains aspects de la méthodologie de Vauquois et de son équipe.

Les grammaires restent faciles à contrôler, l'organisation hiérarchique permettant de réaliser une stratégie efficace. On peut, par exemple, isoler les grammaires qui traitent les cas compliqués, et ne les utiliser qu'en cas de nécessité. C'est seulement l'un des paramètres laissés à la liberté de l'utilisateur. Disposant du choix des grammaires, de leur organisation en hiérarchie, du statut de ces grammaires et de statut de chaque règle dans chaque grammaire,

on peut imposer à l'algorithme son choix de construction de l'arborescence transformée.

L'organisation séquentielle des modèles, propre aux travaux de CETA, a été remplacée par des systèmes intégrés où l'utilisateur n'est plus prisonnier d'une seule stratégie. L'innovation du système vient du fait qu'il donne au linguiste non seulement une vaste gamme de modèles, mais qu'il lui offre également une considérable liberté dans la stratégie et dans l'exploitation de l'heuristique à insérer dans le processus.

Les débuts d'utilisation effective des techniques mises au point en laboratoire font apparaître un certain nombre de problèmes intéressants. Un des principaux avantages du type de représentation choisie est celui d'être strictement arborescent, d'où vient sa clarté. Néanmoins, il n'est pas toujours évident qu'il soit assez complet pour contenir tous les énoncés d'une langue naturelle. Nous pourrions bien nous demander si les énoncés d'une langue naturelle peuvent se représenter de façon arborescente.[74] Là encore, il semble que l'on doive faire un compromis entre la qualité de la représentation choisie et le coût de l'analyse.

---

[74] Une certaine faiblesse de ce moyen d'expression se révèle lors du traitement des cas d'anaphore complexes, par exemple: "La lune est dans le ciel; pour le voir...". Il est toutefois possible de trouver un énoncé équivalent dont la représentation est aisée à construire. Ici: "Pour voir que la lune est dans le ciel...". Le problème ne semble pas être l'existence d'une représentation pivot, mais sa construction.

Le descripteur utilisé est une structure multiniveau, à partir de laquelle est engendrée la structure profonde. Guilbaud constate : "L'intérêt d'une structure profonde réside dans la possibilité de paraphrasage." [75] La même structure profonde peut-elle représenter plusieurs phrases synonymes non seulement dans la langue de départ mais aussi dans la langue d'arrivée ? A moins d'atteindre un niveau d'universalité hors de portée en ce moment, il peut y avoir des phrases synonymes qui ne sont pas représentées par la même structure profonde. La reconnaissance des structures équivalentes est nécessaire pour la phase de transfert et parfois aussi pour la phase de génération.

Le développement d'un système de TAO, surtout dans un cadre multilingue, exige beaucoup de temps et implique la collaboration de nombreux spécialistes : les linguistes et les lexicographes jouant un rôle plus important que celui des informaticiens. L'utilisateur doit être compétent en matière des logiciels (les LSPL et l'environnement), et pour cela il faut y consacrer un travail de longue durée. Le procédé systématique qui consiste à accumuler les propriétés attribuables à toutes les unités lexicales, est un travail long et patient qu'il est impossible aux chercheurs de réaliser de manière exhaustive sur toutes les langues dont ils

---

[75] Guilbaud. Descripteur Linguistique, 4.

s'occupent. Donc, le GETA veut échanger les données lexicales importantes préparées dans d'autres centres contre les outils informatiques qu'ont développés les chercheurs de GETA. Ils aimeront procèder à des échanges réguliers avec les équipes suivantes: Université de Paris VI - Groupe de D. HERAULT, Université de Montréal - Projet TAUM, Université de Pise - CNUCE, Université de Saarbrücken - Romanistisches Institut, et Université des Langues étrangères de Moscou (Prof. ROSENTZWEIG et Prof. Mel'CHUK). De cette façon, le GETA serait en mesure d'acquérir des dictionnaires exceptionnellement riches en échange de leurs logiciels. Quant à la maintenance, nous observons combien il est facile d'augmenter rapidement la taille d'un dictionnaire, mais difficile par la suite de le corriger.

4.5 PROJET EUROTRA

Le développement d'ARIANE se prolonge dans le cadre d'un projet international. En mars 1977, le GETA a organisé un colloque Université-Industrie avec la participation des autorités d'administration intéressées (y compris la Communauté Économique Européenne). Vauquois en précisait les visées :

> sensibiliser les autorités compétentes et les promoteurs éventuels à cette activité dont le débouché est assuré, pour montrer ce que le laboratoire avait

déjà réalisé, et pour amorcer la phase de développement industriel.[76]

Donc, le GETA a également participé très activement à l'élaboration du projet Eurotra.[77] L'environnement de programmation d'Ariane a contribué à la définition de l'architecture informatique et linguistique de ce projet.[78] Pourtant, depuis 1982, le GETA, pris par le PN-TAO (Projet National de TAO), cesse de jouer un rôle très actif dans la préparation d'EUROTRA. C'est le Centre d'Études Linguistiques pour la Traduction Automatique (CELTA) de Nancy qui devient le principal partenaire français.

Le lancement d'EUROTRA par la CEE en décembre 1982 surprend tout le monde, car les projets préparatoires s'éternisaient. Puisque aucune équipe n'avait d'expérience quant aux langages spécialisés et aux environnements logiciels pour la TAO, deux contrats furent confiés aux équipes du GETA. Appréciant les complications inhérentes au projet, les

---

[76] Bernard Vauquois. "L'Informatique au service de la traduction", dans ANALECTES, 505.

[77] Le traité de Rome accorde une égalité aux langues officielles de tous les États-Membres de la Communauté Européenne. La tâche de traduire tous les documents utilisés ou publiés constitue un véritable défi. Eurotra est donc un projet plurilingue de traduction automatique entrepris conjointement par la Commission et les États-Membres.

[78] Ch. Boitet. "The French National MT-Project : Technical Organization and Translation Results of CALLIOPE-AERO." Computers and Translation. 1:4 (1985): 239-267.

Grenoblois cherchent sans cesse à clarifier la situation. Pourtant, certaines difficultés d'ordre organisationnel et financier persistent. Vers mai 1985 l'EUROTRA-F se met en place et en 1986 Boitet apprend que cet organisme a déjà des problèmes majeurs : "Il semble que certains laboratoires concernés n'ont vu dans ce projet qu'une source commode de financement, sans montrer aucun intérêt pour la TAO, à laquelle ils ne s'étaient d'ailleurs jamais intéressés." [79]

Pour rattraper le retard pris par EUROTRA-F, les chercheurs du GETA proposent d'utiliser le générateur du français du GETA, et l'analyseur du français de B'VITAL[80] (développé lors du PN-TAO). En 1987 les dates limites ne sont pas encore respectées, mais le GETA continue dans la mesure de ses moyens. En l'absence du logiciel EUROTRA, le GETA réalise en Ariane un transfert des structures produites par l'analyseur de B'VITAL.

---

[79] Ch. Boitet et al. Le point sur la participation du GETA à EUROTRA. Document interne, Grenoble, 1988, 2.

[80] B'VITAL (Bernard Vauquois Informatique et Traitement Automatisé des Langues) Fondé en 1985, ce groupe se spécialise dans des applications langagières et dans le développement d'outils de gestion. Ils conçoivent et développent des usages pratiques d'Ariane. Par exemple, ils ont développé un système BDTAO (Bordereaux et Dictionnaires pour la TAO) pour faciliter l'accès aux dictionnaires d'Ariane.

4.6 PROJET ESOPE ET PROJET NATIONAL DE TAO

Devant l'exemple japonais,[81] le GETA commence à trouver nécessaire d'automatiser la traduction de gros volumes de documentation. Comme l'application russe-français permettait d'examiner le système de TA du veilleur, des applications en milieu industriel permettraient de valider les recherches en TA du réviseur. Ariane-78 a été choisi pour la mise en oeuvre d'un prototype industriel (français-anglais) et dans le cadre du projet national de TAO placé sous l'égide de l'Agence de l'Informatique. Grâce au soutien de la Direction des Industries Électroniques et de l'Informatique (DIELI), puis de l'Agence pour le Développement de l'Informatique (ADI), un transfert technologique vers l'industrie a été organisé, sous forme du projet pilote de TAO dans le cadre du projet ESOPE (82-83), puis du projet national de TAO (83-87).[82]

Le GETA a été l'animateur du projet ESOPE de l'ADI, dont l'objectif était de préparer un "projet mobilisateur de la filière électronique", le "PN-TAO" (Projet National de TAO). Le pré-projet ESOPE, mené en coopération avec CAP-Sogeti, a

---

[81] Voir T. Tsutsumi, et al. "Example-Based Approach to Machine Translation". Proc. Premières journées franco-japonaises sur la traduction assistée par ordinateur. Ambassade de France au Japon, Tokyo, Japon. 15-16 mars 1993, vol 1/1, 1993, 161-169.

[82] F. Peccoud. "The Aims of the French National Project of Computer-Aided Translation." International Forum on Information and Documentation. 13/1, 1988. 11-13.

développé une spécification externe complète du système Ariane-78.3 et une méthodologie industrielle de génie logiciel,[83] un système de visualisation des données lexicales de tout ou partie d'une application de TAO sous Ariane, y compris la spécification et la mise au point de deux maquettes pédagogiques anglais-français et français-anglais.[84]

On a ainsi réalisé les problèmes d'ingénierie linguicielle que posaient l'écriture de gros systèmes transformationnels. Vauquois et Chappuy inventent un nouveau formalisme, dit de la "grammaire statique". La spécification "statique" des grammaires "dynamiques" est un langage qui vise

---

[83] Nous avons recours au terme de Boitet. "Génie linguiciel": "des techniques de construction et de maintenance de grandes bases lexicales et grammaticales". Ch. Boitet. "La TAO à Grenoble en 1990 : 1989- : vers la TAO du rédacteur (TAO personnelle)." La TAO à Grenoble en 1990. Grenoble: IMAG, 1990, 9.
Le génie logiciel se compose des techniques pour faire des logiciels (les systèmes et les programmes informatiques). Le génie linguiciel comprend les applications linguistiques faites à partir de logiciels spécialisés rédigées par les linguistes. Ces derniers peuvent définir les grammaires et les dictionnaires. Il y a donc trois niveaux de travail: les logiciels écrits par les informatiques, les linguiciels ("programmes") faites par les linguistes, et la mise en oeuvre des programmes par le linguiste.

[84] Nous renvoyons le lecteur à B. Vauquois. "Traduction assistée par ordinateur. Formation de spécialistes : Préparation du transfert technologique." Projet ESOPE, Contrat ADI, mai 1981.
B. Vauquois et al. "Définition d'une méthode de travail d'équipe linguistique". Projet ESOPE, Contrat ADI, novembre 1982.

à la formulation de modèles linguistiques.[85] Elle représente un modèle linguistique et sert comme référence pour tous les programmes d'analyse et de génération conçus pour calculer ce modèle. A partir d'une grammaire statique donnée, l'on peut écrire des programmes (dynamiques) d'analyse ou de génération différents, incorporant des stratégies diverses. Une fois qu'un programme (une grammaire dynamique) atteint un volume certain, une grammaire devient absolument nécessaire pour permettre à quelqu'un qui n'a pas fait partie du développement du programme d'y avoir accès. En outre, une telle grammaire est essentielle au développement modulaire, lorsqu'une équipe collabore de façon autonome à l'écriture des modules d'une grammaire dynamique. C'est le cas du développement industriel d'un système de traduction. "La nécessité d'une spécification "statique" des grammaires "dynamiques" d'analyse et de génération apparaît à Vauquois dès 1980." [86] A partir de grammaires statiques, l'équipe a défini une méthodologie de spécification et d'implémentation d'analyseurs et de générateurs durant le projet ESOPE (à l'occasion de la construction d'une maquette pédagogique anglais-français BEX-

---

[85] "Il faut dire qu'on a là l'exemple d'un formalisme manifestement utile, muni d'une sémantique "intuitive" assez claire, mais dont une sémantique formelle adéquate a été assez difficile à cerner." Boitet. ANALECTES, 7.

[86] Ch. Boitet. "La TAO à Grenoble en 1990. 1980-1990 : TAO du réviseur et TAO du traducteur." La TAO à Grenoble en 1990. Grenoble: IMAG, 1990, 14.

FEX).[87]

Le but initial du PN-TAO était la réalisation d'un système français-anglais, pour des manuels de maintenance d'avions, avec un poste de travail pour traducteur et lexicographe. Au bout d'un an, cet objectif a été complété par celui de spécifier et développer une version améliorée d'Ariane, dite "Ariane-X". De 1983 à 1987, le GETA a travaillé à la spécification linguistique du système français-anglais pour l'aéronautique, construit avec le concours des partenaires industriels.[88] Puisque ses chercheurs augmentaient les capacités et la fiabilité d'Ariane-78.4, le GETA y a joué un rôle important. Le travail linguistique déjà effectué sur le français-anglais a été poursuivi par la société B'VITAL. Ces recherches ont continué dans les années quatre-vingt-dix dans le cadre d'une action du ministre de l'Industrie, au niveau de SITE, la plus grande société européenne de documentation et de traduction technique, dont B'VITAL est devenue filiale.

En 1989, grâce au soutien du Club Informatique des

---

[87] B. Vauquois et S. Chappuy. "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, August 14-16, 1985.

[88] Ch. Boitet. "The French National MT-Project : Technical Organization and Translation Results of CALLIOPE-AERO." Computers and Translation. 1:4 (1985): 239-267.

Grandes Entreprises Françaises (CIGREF)[89] le système Ariane est en train de passer du stade de l'artisanat à celui d'une véritable application industrielle. Le CIGREF commence à organiser des tables rondes pour examiner la TAO et la traduction éventuelle d'énormes quantités de documents. Ceci coïncide avec la décision du ministre de l'Industrie en 1989 d'appuyer un projet de TALN (Traitement Automatisé des Langues Naturelles) dans le domaine industriel français. SITE, une fusion des sociétés de documentation technique les plus grandes de la France (Sonovision and ITEP[90]), est la compagnie de documentation choisie pour mener l'activité.

La première phase du Projet National (début 1990 - mi 1991) vise à l'incorporation d'Ariane à la chaîne de production documentaire. Une deuxième phase aurait comme but l'industrialisation des textes anglais-français et français-anglais. Ensuite, le système s'étendrait aux autres langues européennes. Dans un avenir à long terme, SITE chercherait d'autres utilisateurs éventuels dans le domaine industriel.

4.7 ARIANE-G5

Au fur et à mesure des versions successives du système,

---

[89] Des directeurs de l'informatique de plusieurs compagnies françaises : Aérospatiale, la Banque Nationale de Paris, Michelin, Dassault et Hachette.

[90] ITEP (Ingénierie Technique et Publicitaire) est une société de service informatique.

certaines faiblesses ont été décelées qui rendent incommodes le traitement de certains phénomènes linguistiques. En 1983 le GETA achève la mise au point d'une cinquième version d'Ariane-78. Remédiera-t-elle à ces insuffisances ?

Des développements considérables ont été apportés au système Ariane, dont la version actuelle, Ariane-G5, est ici brièvement exposée. Plusieurs outils logiciels complémentaires y ont été ajoutés. Les techniques de programmation linguistique en TAO sont en train de passer du stade de l'artisanat à celui d'un logiciel.

Ariane-G5 représente un environnement de programmation destiné aux linguistes créateurs de modèles de TAO,[91] aux traducteurs et aux réviseurs. Il repose sur cinq LSPL (ATEF, EXPANS, ROBRA, SYGMOR et TRACOMPL). Ce système tourne sous VMSP/CMS,[92] sur gros ordinateurs (3090, 303X), sur mini

---

[91] Parmi les applications disponibles il existe trois types de linguiciels (grammaires ou dictionnaires) disponibles à l'heure actuelle : linguiciels élaborés en travail interne, ceux faits en coopération avec des universités étrangères ou en collaboration avec des industriels. Voir J.-Ph. Guilbaud et N. Nedobejkine. Rapport sur Ariane-G5 : Point de vue utilisateur linguiste. Document interne EUROLANG, Grenoble, 27-28 février 1992, 12-16.

[92] Ce sont des systèmes d'exploitation (operating systems) propre à IBM. VMSP est un hyperviseur qui stimule un ensemble de "machines virtuelles". Chaque machine virtuelle tourne sous un système d'exploitation propre. Le sous-système RSCS permet d'organiser les machines virtuelles et les ressources réelles du système en réseau. CMS est un système d'exploitation interactif mono-utilisateur puissant, qui supporte un grand nombre de langues et d'outils de programmation.

(43XX, 937X), et gros micro (PC-AT/370, PS2-80/7437). Ariane-78.4 avait deux versions parallèles, en français et en anglais. Ariane-G5 est programmé de façon totalement multilingue, mais la seule version complète est en français. Par rapport aux systèmes existants, Ariane-G5 présente l'avantage que l'unité de traduction n'est pas réduite à la phrase, mais peut comporter jusqu'à environ deux cents occurrences (mots).

Comme la version précédente, la traduction d'une langue de départ vers une langue d'arrivée se fait en trois étapes successives : analyse, transfert et génération. Chaque étape est réalisée en au moins deux et au plus quatre "phases" successives, selon le schéma ci-dessous. Certaines sont optionnelles (celles dont le sigle termine par X ou Y). Chacune est écrite dans l'un des quatre LSPL supportés par le système, et notés dans des cercles. Lors de l'exécution de l'étape correspondante sur la représentation d'une unité de traduction, le code sert à guider l'interpréteur général du LSPL considéré.

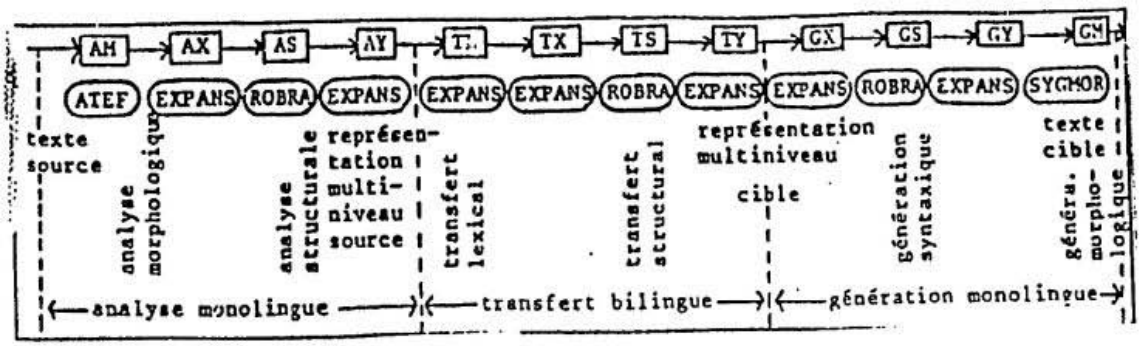


Fig. 4. Dans la version 5 de ce système le schéma d'un processus de traduction automatique est donné par le tableau.

Depuis Ariane-78, de nombreuses extensions ont été apportées, la sémantique précisée, le "moteur" totalement respécifié et réécrit, mais le modèle algorithmique n'a pas varié. ATEF, conçu en 1971 par J. Chauché, est un langage pour l'analyse morphologique. Entre AM, la phase d'analyse morphologique, et GM, celle de la génération morphologique, l'unité de traduction est toujours représentée par un arbre décoré. Les phases écrites en EXPANS sont celles d'expansion lexicale: chaque noeud de l'arbre est remplacé par un ou plusieurs noeuds, selon son unité lexicale et un contexte réduit. Par exemple, l'étape de transfert lexical permet

d'écrire des dictionnaires bilingues multichoix. Si le contexte n'est pas suffisant pour choisir un équivalent, on peut en produire plusieurs et choisir le bon lors d'une étape ultérieure. ROBRA, le successeur du langage CETA, est un langage d'écriture de systèmes transformationnels agissant sur des arbres décorés. SYGMOR est un langage pour la génération morphologique. Il prend en entrée une suite de décorations et produit en sortie une chaîne de caractères.

Plusieurs outils ont été ajoutés à Ariane-G5 :

(i) Normalement le réviseur travaille devant l'écran à une, deux, trois ou quatre fenêtres, dans lesquelles peuvent apparaître le texte de départ, la traduction brute, le texte révisé, et un dictionnaire informatisé.[93] Le GETA a mis au point un système appelé THAM (Traduction humaine aidée par la machine), maintenant intégré à l'environnement REVISION d'ARIANE-78. Ce système enrichit les fonctions classiques d'un éditeur multi-fenêtres (XEDIT) de fonctions spécialisées pour la traduction ou la révision humaine. Il ne s'agit pas d'un poste du traducteur/réviseur complet, mais d'une extension utile à XEDIT. Supposons

---

[93] La révision a un double but: améliorer la qualité linguistique de la traduction et recenser les différents types d'erreur, et les classer selon leur effet sur la révision.

que l'utilisateur soit en train de réviser une traduction automatique. Il peut avoir recours à THAM, qui lui permet d'accéder à un dictionnaire modifiable en sus de la traduction brute et de l'original.

(ii) LT est un langage d'écriture de transcripteurs. Il permet d'écrire rapidement des transcripteurs de textes. Un transcripteur écrit en LT permet d'obtenir les textes en cyrillique avec des majuscules, minuscules et formatage minimal.

(iii) L'évolution lexicale du système consiste à enrichir les dictionnaires d'une façon systématique. Réalisant que ce genre d'opération est non seulement très délicat mais très coûteux, le GETA a mis en oeuvre le programme ATLAS, qui fournit une aide informatisée à l'indexage.

(iv) Cette expérience d'utilisation a amené à définir et à réaliser un environnement spécialisé, PROTRA. Pour la mise au point d'une traduction sous Ariane-G5, toutes les opérations (saisie, vérification, traduction...) peuvent se faire dans le même espace utilisateur. Il est plus convenable d'associer un espace utilisateur à chaque étape fonctionnelle. PROTRA est un environnement qui doit gérer un réseau de machines, pouvant communiquer entre elles. De cette façon, il est possible de partager le module de traduction entre plusieurs machines.

(v) D'ailleurs, des développements importants ont été apportés à l'organisation des dictionnaires. La notion d'unité lexicale est utile en ce qu'elle permet la représentation compacte des familles dérivationnelles. Il est possible de diminuer la taille des dictionnaires et d'aborder les néologismes. Avec Ariane-78, il fallait représenter toute l'information lexicale dans les dictionnaires d'analyse morphologique. Par contre, Ariane-G5 offre des phases d'expansion lexicale qui permettent de répartir l'information. Par exemple, l'analyse morphologique fournit les moyens de passer des morphes (bases, affixes...) aux lemmes, ou l'analyse expansive "X" écrite en EXPANS pour passer aux unités lexicales, et l'analyse expansive "Y" (également écrite en EXPANS) pour traiter les tournures non figées ou non connexes (ex : verbes à particules en allemand).

Il s'agit essentiellement, d'outils de gestion destinés à incorporer les techniques de TAO dans l'activité de traduction. Une expérience de ce type est irremplaçable pour déterminer de quelle façon des techniques d'intelligence artificielle (systèmes visant à une compréhension "explicite") peuvent venir compléter les techniques de "compréhension

implicite" (traitements formels) utilisées jusqu'ici en TAO.[94]

Les trois types de TAO précédents représentent maintenant la TAO "classique".[95] La TAO pour le veilleur des années soixante fournit des traductions "grossières", effectuées rapidement, en grand volume et à bas coût. Celle pour le réviseur, comme Ariane, produit automatiquement des traductions "brutes", destinées à être révisées. Elle n'est envisageable que pour les textes homogènes extensifs, comme des manuels d'utilisation ou de maintenance. Une condition essentielle de succès de ce type de TAO est de constituer une équipe de développement et de maintenance des linguiciels (dictionnaires et grammaires) qui soit en liaison constante avec l'équipe de révision. Ces systèmes doivent ensuite être complètement maîtrisés par leurs utilisateurs, leur donnant la possibilité de les faire évoluer de façon appropriée. Enfin la THAM (Traduction Humaine Assistée par la Machine), selon laquelle l'utilisateur traduit au moyen d'un poste de travail, ouvre sans doute une voie plus réaliste. Des difficultés,

---

[94] Nous empruntons les définitions de Ch. Boitet et N. Nedobejkine. "L'informatique au service de la linguistique : Illustration sur le développement d'un atelier de traduction automatisée." La recherche française par ordinateur en langue et littérature. Actes du colloque, l'Université de Metz, juin 1983. Genève: Slatkine, 1985, 145.

[95] Ch. Boitet. "La TAO à Grenoble : Présentation générale", 2.

tant informatiques que linguistiques, demeurent. Pourtant, il n'en est pas moins vrai que ces procédés sont utilisables, et utilisés dans des cadres opérationnels.

5.0 LE GETA ET LA TA DU RÉDACTEUR

5.1 TRADUCTION DU RÉDACTEUR

Malgré l'expansion des services de traduction, ceux-ci ne peuvent pas répondre aux besoins des organismes du monde commercial. L'idéal serait de rédiger en la langue maternelle et de transmettre les textes tels quels à l'étranger, où la traduction se révélerait nécessaire ou non. Cela impliquerait la valorisation d'une ou de plusieurs langues internationales mais la défense des langues nationales est d'actualité. Ainsi le GETA vise "un moyen concret de concilier la promotion des langues nationales et la nécessité de communiquer en langues étrangères".[96] Pour la communication personnelle, enseigner plus de langues est certes une solution valable. Toutefois, cela ne résoudra guère, dans le cadre européen, le problème de l'auteur d'un article qui voudra écrire dans sa langue maternelle et être compris dans les huit autres. Les anciennes techniques de TAO ne pourront jamais satisfaire à ces nouvelles exigences. La TAO du veilleur, sans préédition ni postédition, ne peut fournir un travail de qualité suffisante. La traduction humaine assistée par la machine (THAM) peut offrir de bons résultats mais il s'agit ici de

---

[96] Ch. Boitet. "La TAO à Grenoble en 1990 : 1989- : vers la TAO du rédacteur (TAO personnelle)." La TAO à Grenoble en 1990. Grenoble: IMAG, 1990, 1.

traduction assistée, et non plus automatique. Enfin, la TAO du réviseur ne s'adresse qu'à des spécialistes au moins bilingues, et non à la plupart des rédacteurs qui sont unilingues. Les chercheurs comptent donc proposer des solutions concrètes aux nouvelles demandes de transformation multilingue de documents scientifiques et techniques de taille moyenne.

Il s'agit de TAO "grand public", de qualité, et en général sans révision (puisqu'on ne peut pas mettre un réviseur à côté de chaque rédacteur). Si c'est le grand public qu'on vise, il faut concevoir un système pour des non-spécialistes. On n'aura pas recours à une base de connaissances pour "comprendre" les textes. On ne va pas non plus se restreindre à un langage "contrôlé".[97] Pourtant, l'utilisateur acceptera "d'uniformiser" le texte à traduire (sous ses aspects lexicaux, grammaticaux et stylistiques) : un "langage guidé". De plus, afin de réduire les ambiguïtés (lexicales, grammaticales et sémantiques), il acceptera de "clarifier" le texte : "interaction avec l'auteur".

Il est possible de concevoir la liaison d'un système de traduction personnelle à un système expert du domaine traité.

---

[97] D'après Boitet "Un des buts du dialogue est alors de standardiser le texte, c'est-à-dire de guider l'utilisateur vers une formulation conforme à ce qui est attendu, tout en le laissant libre, pour un champ donné, de modifier le type de fragment prévu." "La TAO à Grenoble en 1990 : 1989- : vers", 7.

Cependant, puisqu'on n'est pas encore arrivé à coder la connaissance nécessaire à une compréhension parfaite du texte à traduire,[98] le nouveau système fait appel à l'intelligence de l'auteur. On aboutira à une compréhension "indirecte", préférable nous semble-t-il, car l'auteur sera toujours plus compétent que tout système expert sur son texte.

D'autre part, il est apparu que la TAO pour rédacteurs, autre sorte de TAO individuelle, pose des problèmes informatiques intéressants dans le domaine du génie logiciel et des interfaces homme-machine. Il faudra considérer une architecture logicielle tout à fait nouvelle. Pour cela les chercheurs devront revoir une bonne partie des méthodes linguistiques, en particulier en ce qui concerne la désambiguïsation, puisque la post-édition directe par un professionnel (le réviseur) devra être remplacée par une préédition indirecte (par l'auteur).

---

[98] Larose insiste sur la relativité du succès de la traduction. D'après lui, il n'existe pas d'équivalences exactes entre les langues. Les théoristes sont en général d'accord pour affirmer que la compréhension parfaite, tout comme la traduction parfaite n'existe pas. Robert Larose. Théories contemporaines de la traduction. 2e éd. Québec: PU de Québec, 1989.

5.2 UTILISATION D'HYPERTEXTE

La grande diffusion des micro-ordinateurs de plus en plus puissants permet d'envisager l'idée de mettre ce système à portée des auteurs de documentation technique, d'articles scientifiques, ou de livres. En 1988 l'apparition d'HyperCard,[99] un hypertexte[100] disponible gratuitement sur Mackintosh, puissant, programmable et extensible, ainsi que de "piles" de documentation technique,[101] ont suggéré au GETA la possibilité d'une "TAO personnelle" pour rédacteurs monolingues d'hypertextes. L'interactivité inhérente à ce type d'outils permet d'espérer que les utilisateurs accepteraient plus facilement d'interagir avec le système pour standardiser et clarifier leurs textes qu'avec un "texteur" usuel.

---

[99] HyperCard est un environnement de rédaction d'hypertexte qui permet le développement rapide d'interfaces diverses. Voir Guide du langage Hypertalk, Apple Computer, 1988.

[100] Il s'agit d'un hypertexte, non d'un texte : de nombreuses documentations techniques multilingues trop petits ou trop hétérogènes pour être traitées par des systèmes de TAO du réviseur.

[101] Par exemple, Renault a commencé vers 1988 à envoyer à ses concessionnaires européens des CD-ROM contenant la documentation d'une voiture, sous forme de piles HyperCard, en 9 langues.
Ch. Boitet. "TA et TAO à Grenoble... 32 ans déjà!" T.A.L. (revue semestrielle de l'ATALA). 33:1-2, Spécial Trentenaire (1992): 71.

Le concept de TAO personnelle n'impose pas l'utilisation d'hypertexte. Ce nouveau concept pourrait également s'employer dans le cadre de traitements de texte plus classiques. Le GETA a choisi l'hypertexte, et plus spécifiquement HyperCard, comme base de LIDIA-1 pour des raisons d'ergonomie, de diffusion, et d'intérêt linguistique:

(i) Du point de vue ergonomique, l'hypertexte privilégie l'interaction facile. L'équipe imagine qu'un rédacteur accepterait plus aisément une interaction linguistique sous hypertexte que sous traitement de texte. Au lieu de changer ses habitudes, l'utilisateur reste dans la logique de l'outil.

(ii) La diffusion est là. Dès l'arrivée d'HyperCard, les hypertextes sont sortis des laboratoires. Pour un prix minime, n'importe qui peut réaliser des documentations.

(iii) Du point de vue linguistique, les parties textuelles sont bien isolées, et connexes. Les traitements de texte fournissent un mélange de codes de formatage, de texte, de figures, de formules, le tout présenté de façon non connexe.

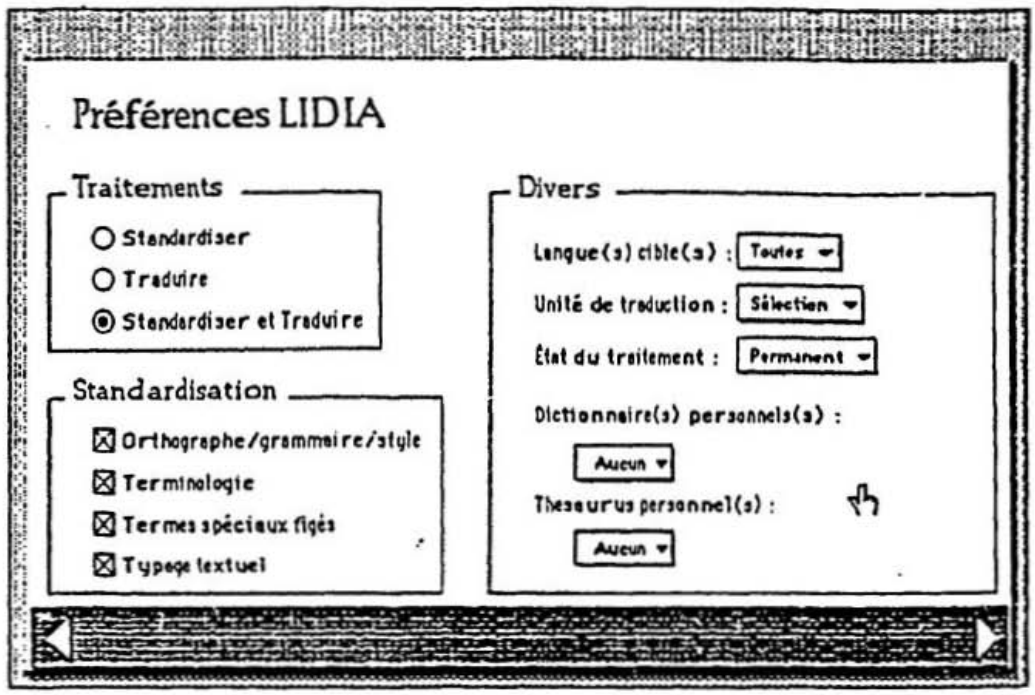


Fig. 5. Sous l'hypertexte de LIDIA-1 l'utilisateur sélectionne entre trois types de préférences.

L'hypertexte est donc un base immensément maniable, et tout à fait adapté à une première expérimentation en TAO personnelle.

5.3 PROJET LIDIA

L'équipe a récemment lancé le projet "Large Internationalisation des Documents par Interaction avec leurs Auteurs" (LIDIA). Dans une première étape, le GETA a construit une maquette de petite taille, LIDIA-1. L'objectif est de simuler d'une manière convaincante les fonctionnalités de traduction du poste de travail du rédacteur, d'aborder certains problèmes linguistiques, informatiques, et ergonomiques, et d'expérimenter diverses architectures logicielles.

Dans la maquette LIDIA-1, seulement deux mille termes sont incorporés, alors qu'un système grand public devrait en comporter quelques centaines de milliers. Les chercheurs se préoccupent des conséquences de l'énormité des bases lexicales et grammaticales qu'un système grand public devrait comprendre pour être viable et prétendent examiner un éventail assez varié de types de textes et de types d'ambiguïtés. Pourtant, ils se limitent à un corpus correspondant à quelques dizaines de pages : deux petites "piles" HyperCard assemblées par J.-Ph. Guilbaud à partir de documents sur Ariane-G5.

L'équipe a également limité les couples de langues recherchés. LIDIA-1 ira du français vers le russe, l'allemand et l'anglais. Boitet explique qu'une même analyse profonde "multilingue" du français peut être utilisée pour la traduction éventuelle vers toutes les langues : "Si on

l'enrichit par l'ajout d'acceptions interlinguales (pour le groupe de langues considéré), on ne peut qu'améliorer les résultats".[102]

Les chercheurs ont opté pour une maquette très réduite devant permettre de valider le concept et d'aborder des nouvelles difficultés qui se présentent en construisant un petit prototype. Ils visent la traduction en partant d'un hypertexte et non d'un texte. Le nombre des "styles" et des "formes" est donc fortement limité. Il est quand même plus facile de parler de "types de fragments" que de "types de documents".

La notion de "sous-langage" a été introduite et étudiée par le linguiste R. Kittredge, directeur du groupe TAUM de l'Université de Montréal au début des années soixante-dix.[103] Sa notion de "clôture lexicale" servait comme définition d'un sous-langage. Afin d'obtenir un cadre adapté à la TAO personnelle le GETA simplifie l'approche de Kittredge, en proposant une définition formelle relativement élémentaire des

---

[102] Ch. Boitet. "La TAO à Grenoble en 1990 : 1989- : vers", 4.
Voir aussi Ch. Boitet. Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation. New Directions in Machine Translation. BSO congress, Budapest, August 1988.

[103] Kittredge voulait déterminer si l'approche deuxième génération avec sous-langage était prometteuse. R. Kittredge. "Sublangage - Specific Computer Aids to Translation: A Survey of the Most Promising Application Areas." Contract no 2-5273, Université de Montréal et Bureau des Traductions, mars 1983.

notions de "micro-langage" et de "sous-langage". Au lieu de se concentrer sur l'aspect lexical, le GETA s'attache à l'aspect grammatical.

L'utilité de ces définitions reste dans la possibilité d'associer à chaque champ le type de fragment (micro-langage ou sous-langage) qu'il est susceptible de contenir. L'interaction avec l'auteur permet de standardiser le texte. Au moyen du "dialogue", l'utilisateur est guidé vers des formulations conformes à ce qui est attendu. Ce dernier est toujours libre, pour un champ donné, de modifier le type de fragment prévu. On peut mettre ce concept en comparaison avec celui de "langage contrôlé". Pourtant, sous LIDIA le texte n'est ni libre, ni contrôlé, ni sous-langage au sens de Kittredge. Sous cette maquette le texte est "négociable". L'utilisateur se sert d'une multiplicité de langages contrôlés, organisés à deux niveaux, celui des énoncés (micro-langages) et celui des textes (sous-langages).

Du point de vue grammatical, le GETA propose une voie intermédiaire, dite du "langage guidé". C'est-à-dire que l'utilisateur se sert de tout un ensemble de langages contrôlés, chacun étant défini soit comme un "style", soit comme une "forme". Selon Boitet, un style est un type d'énoncé, et une forme est un type de texte, lui-même défini à partir de styles et de formes par une grammaire hors-

contexte étendue.[104]

Puisque le GETA vise une diffusion auprès d'un large public, l'équipe désire que la station de rédaction soit un micro-ordinateur convivial et largement diffusé, d'où le choix du Macintosh sous HyperCard. En outre, il désire réutiliser une fois de plus la puissance de leur générateur de systèmes de TAO (Ariane-G5). Il veut s'en servi pour rédiger les parties accablantes du traitement linguistique. Pourtant, Ariane-G5 n'est pas encore disponible sur Macintosh. De plus, Vauquois constate que l'exécution d'Ariane-G5 est toujours trop lente sur ce type de matériel, et augmenterait sûrement les temps de réponse.[105] L'équipe a donc choisi un traitement distribué entre la station de rédacteur et un serveur, et un fonctionnement asynchrone.

Influencé par l'approche du système CRITIQUE d'IBM,[106] le GETA a opté pour une architecture distribuée, comportant un réseau de stations de rédaction (Macintosh) sur lesquelles on effectue les opérations linguistiques légères et guide le dialogue, couplé avec un serveur de TAO sur mini-ordinateur

---

[104] Boitet. TA et TAO à Grenoble...32 ans déjà!, 73. Dans les documents précédents, le GETA avait proposé "microlangage" pour "style" et "sous-langage" pour "forme".

[105] Boitet. "LA TAO à Grenoble en 1990. 1989-: vers", 7.

[106] Voir S.D. Richardson. Enhanced Text Critiquing using a Natural Language Parser : the CRITIQUE system. IBM Research Report RC 11332, Thomas J. Watson Research Center, Yorktown Heights, 1985.

(IBM-4361). On réalise les phases linguistiques lourdes au moyen de linguiciels écrits en Ariane-G5. L'option de l'architecture distribuée a un base pratique : Le traitement de langue naturelle au moyen des machines petites et bon marché simplifierait à l'extrême les composantes linguistiques et dégraderait la qualité des traductions. Grâce à l'architecture distribuée, le rédacteur n'est pas pénalisé par des résultats inacceptables, des interruptions ou des attentes imposantes.

Sur la station de rédaction, l'utilisateur dispose d'un lemmatiseur, un correcteur orthographique, un dictionnaire des sigles, un dictionnaire des tournures figées, un thésaurus, un dictionnaire des sens muni de questions permettant de choisir entre des mots ambiguës (e.g. river/rivière et river/fleuve), et éventuellement un analyseur syntaxique superficiel permettant de détecter des fautes d'accord. De plus, chaque pile à traduire s'accompagne des résultats des divers traitements.

Le serveur de TAO comporte des dictionnaires et des grammaires de systèmes de TAO "classiques" rendant possible l'analyse morphologique ; une première analyse syntaxique exposant les diverses ambiguïtés insolubles sans recours au rédacteur; une seconde analyse, plus profonde, par interaction avec le rédacteur; un ou plusieurs transferts vers les langues d'arrivée ; et les générateurs correspondants.

Examinons le processus linguistique. Dans l'espoir de simplifier le processus d'interaction lors de la phase de standardisation l'orthographe du texte est contrôlée, les tournures figées exposées et une forme est désignée au texte. Le serveur de TAO analyse ce texte et le transmet à la station de rédaction. Par la suite, une phase de dialogue avec le rédacteur permet d'obtenir la structure "multiniveau" désambiguisée de l'unité de traduction. Celle-ci est "réduite" à une structure multiniveau abstraite. A l'aide des composants de transfert le système fournit les structures multiniveaux abstraites d'arrivée. Le processus de traduction se termine par la génération structurale, syntaxique et morphologique. Dès lors, afin de contrôler le résultat, le rédacteur (monolingue) peut exécuter une rétrotraduction en se servant de la formule d'arrivée.

Parmi les aspects les plus intéressants de ce nouveau concept certaines possibilités d'utilisation sont à noter : la synthèse vocale dans le dialogue (intégration de synthèse vocale, dialogue de clarification, sortie de traductions), l'utilisation de rétrotraductions (la vérification indirecte des traductions), et la possibilité d'apprentissage assisté des langues (par accès aux traductions et aux dictionnaires supportés par la station de rédaction).

Parmi les problèmes actuels de recherche :

(i) La représentation des structures soumises à un programme. L'essentiel c'est de concevoir une structure de données adéquate pour représenter une pile ou un document. D'abord, la transcription d'entrée, sous laquelle les unités de traduction sont envoyées au serveur de TAO, doit permettre de coder tous les résultats de la préédition indirecte effectuée grâce au dialogue (choix de sens, des classes, des types de fragment, etc.) Ensuite, les représentations arborescentes fournies aux différentes étapes doivent être adaptées pour faciliter la construction du dialogue de clarification.

(ii) La représentation des connaissances linguistiques. Puisque l'utilisateur n'est pas censé être grammairien, la représentation des microlangages et sous-langages est délicate.

(iii) Il faut donner aux utilisateurs la possibilité de faire évoluer eux-mêmes ces connaissances, par exemple en ajoutant des néologismes (y compris des termes traditionnels avec changement de sens), ou bien de nouveaux sous-langages.

(iv) Enfin, le traitement lui-même est problématique. Il faut le distribuer entre des micros reliés sur un réseau réel et des "machines virtuelles" organisées en "réseau virtuel" (sous VMSP.5 sur IBM-4361).

La maquette envisagée ne comporte que quelques milliers de termes au maximum, alors qu'un système grand public devrait en comporter quelques centaines de milliers. En arrivant à un système opérationnel destiné au grand public il est bien probable que les chercheurs vont se heurter à des obstacles non-prévus ou sous-estimés, du point de vue ergonomique, économique et technique. Le GETA s'impose de prendre en compte la richesse lexicale des termes par rapport à toute la langue, sans la réduire aux acceptions attestées dans les piles de test. En effet, des simplifications abusives pourraient conduire à des solutions simplistes et erronées.

Les problèmes ergonomiques seront également de grand intérêt. Il faut passer un temps considérable à rajouter au système Ariane-G5 les outils spécifiques de LIDIA. Il faudra aussi refaire des outils logiciels, de manière à obtenir une version homogène, portable sous tous les environnements (OS/2 avec PM ou Windows, Unix avec Motif, Mac-OS, etc.). Enfin, un tel effort ne pourra se justifier que si la maquette donne de bonnes raisons d'espérer qu'on peut arriver à un système utilisable par un large public.

Si le GETA veut proposer des solutions pour répondre aux besoins du rédacteur qui désire rédiger dans sa langue, et transmettre à l'étranger, il faut tenir compte des problèmes de la TAO classique. Ces problèmes, tant linguistiques qu'informatiques se posent toujours. Certains sont plus

aigus: la levée d'ambiguïtés (au moyen du dialogue de clarification), le guidage du rédacteur vers des micro-langages ou des sous-langages (au moyen du dialogue de standardisation), etc.

Pour que la traduction de rédacteur débouche sur des solutions utilisables, il faudra aussi sans doute que de nombreux groupes de recherche s'engagent dans cette voie, et que les investissements suivent. Pour passer d'une maquette comme LIDIA à un produit, il faudra trouver la solution de nombreux problèmes. Selon Boitet, le plus important est celui du "génie linguiciel".[107] Le GETA sera obligé de développer au moins la partie lexicale par des efforts coordonnés des groupes de recherche en traitement des langues, ainsi que des industriels du logiciel et des fabricants de dictionnaires usuels.

5.4 BASE LEXICALE NADIA

Le projet LIDIA nécessite, pour passer le stade du prototype, une base lexicale de très grande taille vérifiant les propriétés nommées. Cette base doit comprendre un grand nombre d'informations lexicales allant de la phonétique à la sémantique. Comme nous l'avons déjà indiqué, l'auteur du

---

[107] "c'est-à-dire des techniques de construction et de maintenance de grandes bases lexicales et grammaticales." Boitet. "La TAO à Grenoble en 1990. 1989- : vers", 9.

document coopère avec le système pour désambiguiser lorsque ce dernier ne peut produire une seule "interprétation" d'une unité de traduction. Le rédacteur choisit une des diverses interprétations qui lui sont présentées (qui correspond "sémantiquement" à ce qu'il a voulu exprimer) au moyen d'une phase de dialogue. Les dictionnaires doivent donc contenir la représentation des différents sens possibles des mots. De plus, pour chaque sens, il faut disposer d'une définition ou d'un synonyme, qui seront proposés au rédacteur.

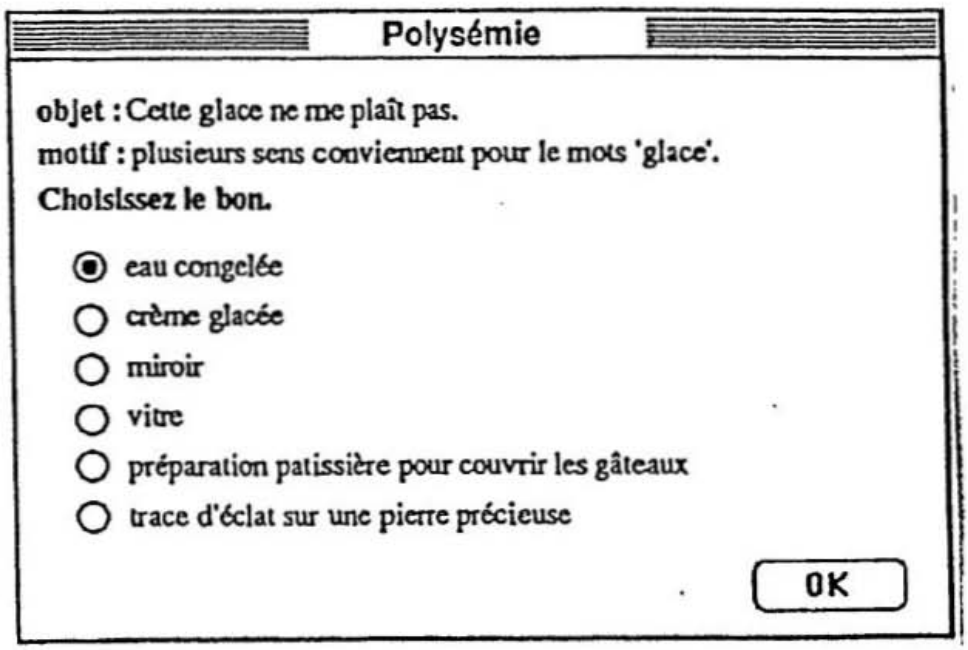


Fig. 6. Lorsqu'une polysémie n'a pas été résolue automatiquement, un dialogue est généré.

Ariane-G5, comme nous l'avons indiqué, impose trois étapes : analyse, transfert et génération. Les données concernant un même lemme[108] sont dispersées dans les dictionnaires propres à chacune des trois étapes. Pour des raisons d'économie de place, les informations dans ces dictionnaires sont représentées sous forme codée. C'est donc seulement le linguiste qui sera en mesure d'effectuer l'indexage. De plus, les informations représentées dans le dictionnaire sont celles qui sont utiles à l'application et ne concernent que cette dernière. Ces dictionnaires ne peuvent donc être réutilisés pour une autre application ou par un autre système.

Or, les besoins en ressources lexicales de grande taille pour les systèmes de TAO augmentent chaque jour. Ces ressources représentent sûrement la partie la plus coûteuse d'un système de TAO. En même temps se révèle une demande croissante pour le développement de dictionnaires réutilisables. Dès 1982, des bases de données lexicales indépendantes de l'outil sont apparues ailleurs, par exemple le projet MU (Japon).

L'indexage d'une base lexicale est long et coûteux. Pour y remédier, des outils d'acquisition automatique de

---

[108] Un lemme est une famille d'occurrences générées par le même paradigme à partir d'une racine généralement commune et appartenant à la même classe. Ainsi, "a, aurons et ont" sont des occurrences du lemme "avoir".

connaissances ont été mis au point.[109] De plus, le GETA prévoit un échange d'informations vers l'extérieur. Pour cela, les chercheurs utiliseront le format SGML (Standard Generalized Markup Language) et suivront les spécifications de TEI (Text Encoding Initiative).[110] La communication avec les autres groupes se fera, par exemple, par la génération de dictionnaires d'applications spécifiques, à partir de la base générique. Serasset envisage un "langage de communication" qui permettra de générer des informations sous le format SGML. Ce mode de communication devra aussi rendre possible une récupération d'informations extérieures. Pour remplir leur base lexicale, ils pourront utiliser les informations d'autres bases, de dictionnaires papier ou encore de dictionnaires d'application. On réduira ainsi le coût de rassemblement de manière significative.

Les dictionnaires que comprennent les systèmes de TAO s'appuient le plus souvent sur l'architecture du système et sont en général peu lisibles par un utilisateur non-spécialiste. Le nouveau projet du GETA vise le développement d'une nouvelle architecture lexicale pour les dictionnaires multilingues. L'avantage de la base lexicale projetée réside

---

[109] La société B'VITAL dispose de dictionnaires d'applications linguistiques construits sous l'environnement ARIANE-G5.

[110] TEI : Guidelines for the encoding and the interchange of machine-readable texts. ACH-ACL-ATC., juillet 1990.

dans sa réutilisabilité.

Cette base lexicale doit non seulement pouvoir servir aux différentes applications de TALN, mais aussi à un utilisateur non qualifié. Les informations seront à la fois utilisables par la machine et compréhensibles à l'utilisateur. De tels emplois supposent une interface homme-machine permettant un accès rapide et efficace aux informations que l'on veut extraire.

Chaque application nécessite des informations de nature différente. Dans un système de TAO, il faut des informations grammaticales et morphologiques, alors qu'un système de synthèse de la parole implique des informations phonétiques. La base de ce projet doit donc contenir des informations fort diverses.

Enfin, le GETA propose une approche interlingue basée sur les acceptions (les différents sens de chaque mot) pour développer son système de gestion de bases de données lexicales. Le but n'est pas de représenter la sémantique de ces acceptions de manière formelle. L'objectif du projet est de disposer de liens structurés entre les langues, chaque lien dénotant l'existence d'un concept. Selon l'approche, le lexique interlingue est l'union des acceptions des langues qui apparaissent dans la base. (voir fig. 7)

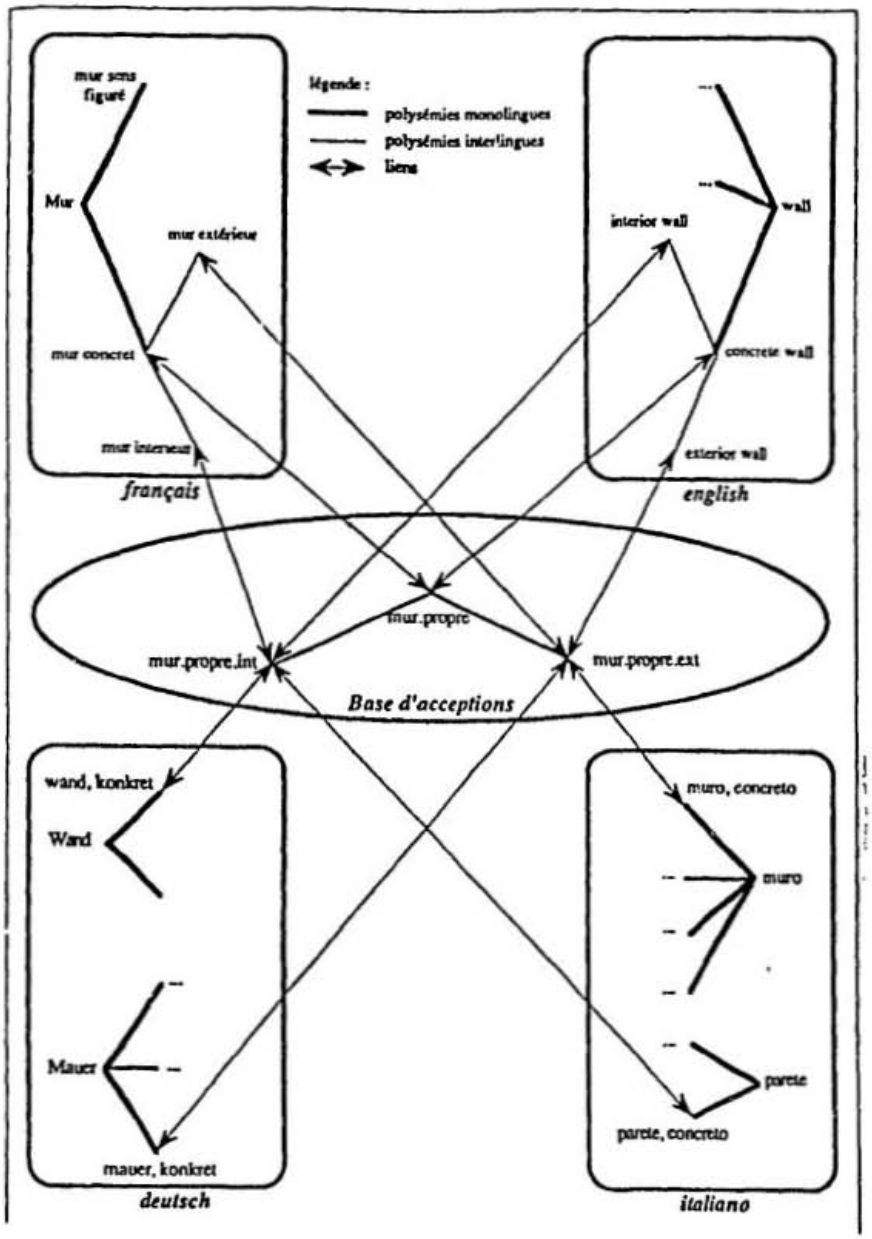


Fig. 7. Exemple d'acceptions et de liens pour quatre langues.

L'équipe recherche la construction et l'utilisation de bases de données lexicales multilingues, indépendantes d'une théorie linguistique et indépendantes d'une application. Pour cela, les chercheurs ont choisi de développer un système de gestion de bases de données lexicales multilingues fondé sur une approche interlingue : NADIA (Nouvelle Approche de Dictionnaires Interlingues par Acceptions). L'unité interlingue sera "l'acception". Au lieu d'adopter l'approche par transfert, où les liens entre les langues se font par l'intermédiaire de dictionnaires bilingues et unidirectionnels (comme Acquilex et Multilex), les liens entre les langues se feront au moyen d'un dictionnaire interlingue unique (comme KBMT-89 - Knowledge-Based Machine Translation[111] et EDR - Electronic Dictionary Research[112]).

La base NADIA devra respecter trois contraintes, imposées tant par le contexte de développement (LIDIA) que par les objectifs que le GETA s'est fixés (utilisation mixte, échange d'informations). Le système est conçu pour être multilingue

---

[111] Voir I. Meyer, B. Onyshkevych et L. Carlson. "Lexicographic Principles and Design for Knowledge-Based Machine Translation", Carnegie Mellon University, Technical Report no CMU-CMT-90-118, 13 août 1990.

[112] Voir "Electronic Dictionary Project", Japan Electronic Dictionary Research Institute, novembre 1988.
"EDR Technical Reports : An Overview of the Electronic Dictionaries", EDR, Japan Electronic Dictionaries Research Institute, Technical Reports nos TR-024, TR-025, TR-026, TR-027, TR-029.

(le système gère des bases multilingues), indépendant vis-à-vis des applications (les bases peuvent être utilisées dans différentes visées : traduction, correction, apprentissage des langues par l'homme), générique (les structures linguistiques utilisées pour les bases seront définies par un linguiste) et indépendant de la théorie (de nombreuses théories linguistiques peuvent mettre en oeuvre de nombreux formalismes informatiques).

L'équipe a adopté une approche interlingue car elle pense apporter la meilleure solution au critère d'indépendance vis-à-vis des applications. Tout en étant compatible avec des applications interlingues, le système peut générer des dictionnaires bilingues à partir de ces bases. La visée est de générer de manière automatique des dictionnaires pour diverses applications linguistiques. Pourtant, si NADIA doit être indépendante des applications spécifiques qu'elle rend possibles, le GETA tient toujours compte des applications éventuelles. Ce système est conçu dans le cadre du projet LIDIA. En conséquence, la base lexicale NADIA devra contenir au moins les informations nécessaires au projet LIDIA.

L'architecture de NADIA est basée sur un dictionnaire pivot multilingue. Les dictionnaires sont organisés "en étoile". Les dictionnaires monolingues se trouvent en périphérie et le dictionnaire interlingue "pivot" au centre. Ce dernier contient les acceptions des différentes langues de

la base. Les dictionnaires monolingues contiennent l'information linguistique des différentes entrées lexicales.

Un dictionnaire monolingue est divisé en deux parties. La première regroupe les acceptions de la langue du dictionnaire (partie purement monolingue). La seconde regroupe les acceptions d'autres langues qui n'ont pas d'équivalent dans la langue du dictionnaire (partie contrastive). Cette dernière vise à représenter les phénomènes de polysémie introduite par le multilinguisme, qui doivent être distinguées pour chaque langue. Puisqu'elles ne sont introduites que par la présence d'autres langues, elles ne sont pas dans les dictionnaires purement monolingues.

Il existe donc deux sortes d'acceptions. La première type regroupe les acceptions monolingues propres à une langue qui doivent être présentées indépendamment des autres langues de la base. Le deuxième type d'acception regroupe les acceptions interlingues dues aux phénomènes contrastifs entre les langues de la base. Ainsi, dans le cas du mot "river", le dictionnaire monolingue anglais ne doit contenir que les deux acceptions correspondant au sens propre et au sens figuré de ce lemme.

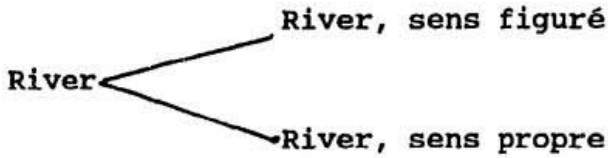


Fig. 8. Dictionnaire monolingue anglais pris indépendamment des autres langues, article "river".

Par contre, si ce dictionnaire est présent dans la base en même temps que le dictionnaire français, il devra contenir un raffinement de sens supplémentaire dans le cas du "river" sens propre. Les dictionnaires monolingues sont séparés en deux parties :

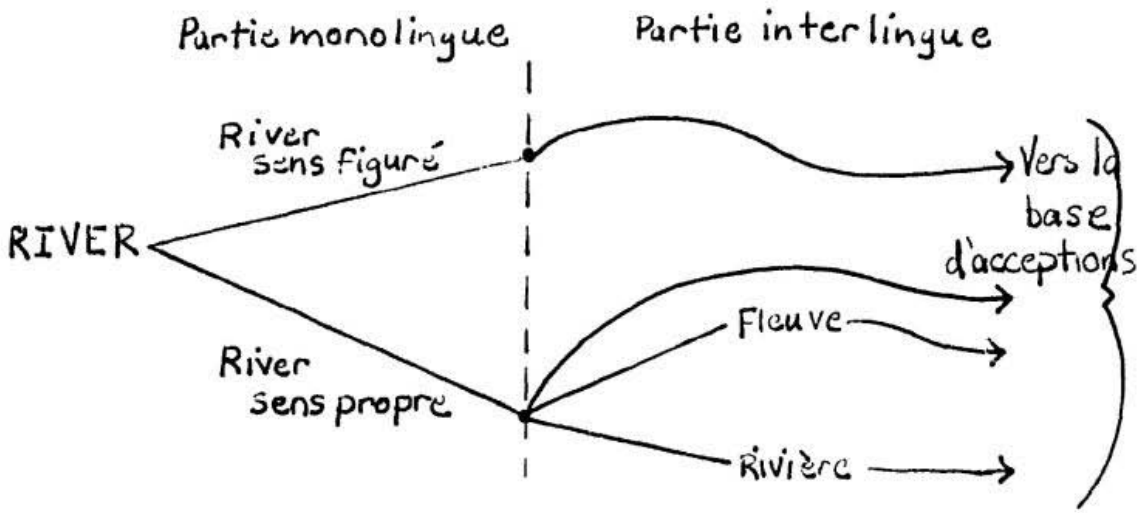


Fig. 9. Dictionnaire monolingue anglais, mot "river".

Puisque l'architecture est fondée sur une base pivot, rajouter une langue à la base lexicale revient à rajouter un dictionnaire monolingue et à mettre à jour le dictionnaire multilingue. Une architecture s'appuyant sur des dictionnaires bilingues imposerait la construction d'un dictionnaire bilingue pour chaque langue considérée, en plus du dictionnaire monolingue. Pour NADIA, des liens seront construits entre les sens des mots du nouveau dictionnaire et la base pivot. De plus, les phénomènes interlingues qui apparaîtront devront être répertoriés sur les dictionnaires déjà présents. Des outils d'indexage seront réalisés pour faciliter le travail des lexicographes.

C'est donc une approche interlingue. Le GETA désigne par "interlangue" un langage artificiel intermédiaire pour réaliser le lien entre les langues. Les énoncés de toutes les langues considérées doivent pouvoir être représentés par cette interlangue. Elle consiste en deux parties distinctes : un lexique et un ensemble d'attributs et de relations.

Le lexique interlingue ou "le dictionnaire interlingues" consiste en l'union ensembliste des acceptions des langues de la base. Ce dictionnaire doit être assez complet pour représenter les différents sens des mots trouvés dans l'ensemble des langues considérées. S'il y a correspondance directe entre deux acceptions de deux langues différentes, celles-ci sont regroupées en une seule acception interlingue.

S'il n'y a pas correspondance directe, les acceptions d'une langue sont conservées dans le dictionnaire interlingue. Quant aux attributs et relations de l'interlangue, l'ensemble n'est pas fixé mais défini pour chaque base par un linguiste.

Le linguiste est libre de définir, pour chaque langue, ses entrées, ses unités lexicales et leurs informations associées, pourvu que le dictionnaire monolingue fasse le lien entre acceptions et entrées. Il définit la structure d'unité lexicale et donne les dérivations qu'il veut prendre en compte. De cette façon, la base lexicale sera indépendante d'une théorie linguistique particulière. L'ensemble des relations et attributs interlingues est aussi défini par le linguiste, à part la relation prédéfinie "isa" (relation de sous-acception), qui est nécessaire dans tous les cas, car les acceptions des termes d'une langue n'ont pas toujours d'équivalent lexical dans une autre langue.

Examinons le processus de traduction. Une interlangue sert à établir un lien entre les langues. Ce lexique interlingue doit fournir un lien lexical entre les mots dans différentes langues. Aussi, deux sens équivalents de différentes langues doivent-ils être reliés à une seule unité interlingue. Lorsqu'il faut choisir le vocabulaire à utiliser dans la langue de départ il arrive que l'on ne trouve pas d'équivalent pour un mot dans la langue d'arrivée. Cela peut se produire lorsqu'une distinction lexicale s'opère dans la

langue d'arrivée qui n'existait pas dans la langue de départ. Par exemple, les mots français "fleuve" et "rivière" sont traduits en anglais par le mot "river". Les deux mots français ont deux sens différents. Dans ce cas un dialogue va être généré afin de trouver le bon équivalent dans la langue d'arrivée. Pour cela, le dictionnaire doit contenir la trace de cette polysémie "contrastive" entre français et anglais. Il faut disposer d'une définition en anglais de cette polysémie introduite par le français. Un lien doit être établi entre ces deux sens dans le lexique interlingue. Cette distinction n'est pertinente que si l'on va de l'anglais vers le français. Dans un contexte de traduction anglais-japonais, cette distinction n'a pas d'importance, puisque le mot japonais "kawa" recouvre le même sens que le mot anglais "river". Aussi, l'interlangue doit-elle contenir une description des différents sens des mots de chaque langue. Des sens équivalents de différentes langues devront avoir des descriptions identiques. Des sens "proches" (comme rivière et fleuve) devront avoir des descriptions "proches".[113]

NADIA demande des études linguistiques considérables. Pour étudier la faisabilité d'une telle approche, le GETA a décidé de construire une maquette : Parax. Elle ne comprend

---

[113] La relation entre acceptions "isa" est prédéfinie. Ainsi, par cette relation, on code que les acceptions de "rivière" et de "fleuve" sont des sous-acceptions de l'acception de "river".

que quelques centaines de mots et comporte pour l'instant 5 langues : français, anglais, allemand, russe et chinois. Pour réaliser la maquette, seule la partie fondamentale du projet a été réalisée, en HyperCard.

Pour la maquette, la structure linguistique des dictionnaires monolingues est inspirée des dictionnaires d'Ariane du GETA. Les raffinements en acceptions strictement monolingues ne sont pas organisés. Une classification systématique représenterait un travail énorme. Aussi, les chercheurs adoptent-ils une structure plate (sans raffinement des sens en sous-sens). En effet, cette maquette n'est qu'une illustration de l'approche interlingue. Elle ne permet aucune indépendance vis-à-vis d'une théorie linguistique. Elle ne fournit pas non plus d'outils spécialisés pour la gestion de bases lexicales multilingues.

Depuis 1990 une nouvelle voie de recherche se dessine. Le GETA veut profiter de NADIA pour stocker des informations permettant d'implémenter des réseaux "neuronaux de désambiguïsation".[114] L'équipe projette de relier entre elles des acceptions pour former un grand réseau. Deux acceptions seront reliées si elles ont un lien sémantique

---

[114] Le principe de ces réseaux est exposé dans J. Véronis, N.M. Ide et S. Harié. Construction automatique de grands réseaux de neurones pour la désambiguïsation du langage naturel, 10èmes journées Systèmes Experts et leur application. AVIGNON'90, Avignon, 28 mai-1 juin 1990, 105-117.

"naturel". Par exemple, les acceptions désignées par "ferme" et "animal" seront reliées. Pour désambiguïser un mot, l'utilisateur fournira un mot et un autre mot important du contexte. Le réseau se stabilisera au bout de quelques instants après avoir "allumé" un chemin entre les deux mots, et après avoir choisi les acceptions voulues.

Ces réseaux seront construits à partir des définitions des dictionnaires. Les chercheurs considèrent qu'un mot est lié sémantiquement aux mots qui composent sa définition. Mais la création d'un tel réseau constitue un travail difficile puisque la construction des liens entre acceptions nécessite le choix des bonnes acceptions pour chacun des mots des définitions et pour cela il faudra désambiguïser les définitions des dictionnaires. De nombreux outils peuvent être mis en oeuvre pour remplir cette tâche. Enfin, le développement d'outils de simulation servira à examiner la faisabilité de cette application à une grande échelle.

Le projet NADIA est la deuxième étape dans le développement d'un système de bases lexicales. Ce projet recherche l'élaboration d'un prototype aussi complet que possible d'un tel système. Le linguiste définit les structures linguistiques des dictionnaires monolingues et gère les informations des dictionnaires monolingues à l'aide d'outils de NADIA (éditeur, défauteur, vérificateur de cohérence). Néanmoins, le dictionnaire interlingue est

difficile à gérer. Une acception interlingue est créée lorsqu'une nouvelle acception apparaît dans une langue. Le lexique doit créer des liens lexicaux entre les différentes langues. Pour cela, les acceptions équivalentes de deux langues différentes doivent être réunies en une seule acception interlingue. Donc, en ajoutant une nouvelle acception, le lexicographe doit vérifier si une acception interlingue équivalente existe dans la base interlingue. Si oui, il va lier cette nouvelle acception interlingue.

Afin de bien gérer ce dictionnaire il faut que le lexicographe puisse vérifier l'existence d'une acception interlingue. Pour ce faire les chercheurs supposent que le lexique contient des définitions d'acceptions qui ne sont pas ambiguës. Puisque le lexicographe doit comprendre cette information, force est de la fournir dans une langue commune partagée par les lexicographes concernés. Pour assurer une cohérence dans la gestion du dictionnaire monolingue, tout en allégeant le travail du lexicographe, le GETA a choisi de confier la gestion de la base interlingue au système. Celui-ci détecte automatiquement les différents problèmes qui peuvent se poser et en propose des solutions au lexicographe.

Le projet NADIA utilise une approche interlingue originale : l'interlingue par acceptions. L'équipe a choisi de donner au linguiste la possibilité de définir ses structures linguistiques. De plus, une structure informatique

de base ne s'impose pas pour le codage des structures linguistiques. C'est ainsi que le système NADIA offre une indépendance vis-à-vis de la théorie linguistique choisie. Le GETA effectue un nouveau pas vers le partage des données linguistiques en permettant à différentes théories linguistiques de cohabiter sur la même plate-forme. Enfin, pour être compatible avec les approches interlingues (EDR et KBMT) et celles par transfert (Multilex), les chercheurs ont choisi une approche interlingue. Celle-ci permet de générer aussi des dictionnaires de transfert.

Alors que les deux projets interlingues ci-dessus utilisent une approche par représentation des connaissances, le GETA croit éviter certains problèmes théoriques et méthodologiques posés par une telle approche. L'approche par acceptions ne nécessite pas un raffinement des unités interlingues (ce raffinement devient systématique, puisque les chercheurs se servent des dictionnaires existants comme référence). Au lieu d'une classification de concepts (comme celle du projet EDR), l'équipe utilise une sorte de relation entre acceptions : une relation de sur-acception à sous-acception. Une hiérarchie d'acceptions ne se désigne pas non plus ; les phénomènes contrastifs sont plutôt codés.

Le prototype de NADIA est en cours de développement. Après son élaboration, il sera possible d'améliorer les différents outils. L'évolution des outils d'importation et

d'exportation rendra plus simple un partage essentiel entre différents systèmes. Les voies de recherche ouvertes par ce projet sont nombreuses : les réseaux "neuromimétiques", les outils d'indexage, etc. Son contexte de développement pourrait bien constituer un projet réaliste et utilisable par des systèmes de TAO ou bien ceux de TALN.

6.0 RECHERCHES ET PROJETS ACTUELS

6.1 MAINTENANCE D'ARIANE ET PROJET EUROLANG

Dès la fin des années quatre-vingt, le GETA prévoit des recherches axées surtout sur le génie linguiciel et sur de nouveaux langages spécialisés.[115] Les nouveaux langages spécialisés permettent de mieux traiter des entrées de lecture optique ou de reconnaissance de parole ; les techniques améliorées de "génie logiciel" pour le "génie linguiciel"[116] permettent d'écrire plus vite des outils informatiques pour linguistes.[117]

Depuis 1990, les chercheurs du GETA se sont limités notamment à sa maintenance quant à Ariane-G5. Ce système est utilisé par le GETA pour le projet LIDIA aussi bien que par des partenaires industriels.[118] Dans le cadre du projet EUREKA EUROLANG, l'équipe vise à la normalisation, à la

---

[115] Boitet. "La TAO à Grenoble en 1990 : 1980-1990 : TAO du réviseur et TAO du traducteur.", 17.

[116] Le "génie logiciel" représente les techniques employés dans la conception des logiciels (les systèmes et les programmes). Le "génie linguiciel" est composé des applications linguistiques faites par les linguistes à partir de logiciels spécialisés.

[117] Voir la thèse de M. Lafourcade. "ODILE-2, un outil pour traducteurs occasionnels sur Macintosh." Colloque "L'environnement traductionnel : La station de travail du traducteur de l'an 2001.", UREF & AUPELF, Mons, avril 1991, PU du Québec, 95-108.

[118] SITE (Sonovision ITEP Technologie) et B'VITAL.

documentation et à l'explication des programmes et des algorithmes d'ARIANE-G5.

EUROLANG est un projet de trois ans qui a été lancé en décembre 1991. Cinq pays européens y participent : la France, l'Allemagne, l'Italie, l'Espagne et l'Angleterre. C'est EUREKA (European Research Coordination Agency) qui est le moteur du projet dont le but initial est de spécifier les éléments informatique et linguistique. L'ensemble sera ensuite développé, intégré et évalué.

Les deux co-organisateurs sont SITE (Sonovision ITEP Technologie) et SIEMENS[119]. Les traducteurs de SITE ont mesuré la qualité des traductions fournies par le système de TAO-ARIANE. Ils sont convaincus que ce système peut être rentable dans un domaine industriel.[120] Pareillement, SIEMENS, qui développe et commercialise le système de TAO METAL[121], montre un vif intérêt pour ce projet. SIEMENS

---

[119] SIEMENS, la société allemande dont la siège sociale est à Munich.

[120] Selon B. Seite le coût absolu d'une traduction au moyen du système ARIANE est actuellement si élevé, que l'avantage procuré dans la réduction du travail du traducteur est diminué lorsqu'on ajoute le coût humain à celui du système.
B. Seite et al. Presentation of the Eurolang Project. Proc. of COLING-92. Nantes, 23-28 août, 1992, 1290.
Voir aussi Bachut, D., et al. "Industrialisation d'un système de TAO français-anglais pour la documentation technique", Génie Linguistique 91, Versailles, 1991.

[121] T. Schneider. "The Metal System", MT Summit III, Washington, 1991.

recherche la définition d'un plate-forme de TALN (Traitement Automatique des Langues Naturelles) commun pour l'Europe et espère améliorer la technologie de son propre système. Son expérience commerciale joue un rôle capital dans le projet EUROLANG.

L'objectif principal de ce dernier est d'assembler une "boîte à outils" pour la TALN, incorporant un grand choix de dispositifs qui reflètent les techniques informatiques et linguistiques les plus récentes. Les composants de la "b-à-o" doivent être destinés aux applications éventuelles du linguiste. Il s'agit de langages spécialisés et de bases lexicales. Les chercheurs procéderont à leur validation grâce à un système de TAO de deuxième génération.

Sur le plan industriel du projet, le système devra permettre aux entrepreneurs européens[122] de mieux communiquer dans un domaine multilingue technique et commercial. Un tel échange est essentiel à l'essor de la coopération internationale.

Afin de développer un système performant et rentable, les partenaires du projet ont décidé de mettre en commun leurs ressources techniques, humaines et commerciales. Pour que le système soit réutilisable et pour garantir la possibilité de son évolution, la plate-forme doit être facile à utiliser et

---

[122] European Business Community

permettre diverses applications de TALN.

Le système sera conçu avec la technologie des systèmes de deuxième génération :

(i) pour séparer les données linguistiques des logiciels les linguistes utiliseront des langages spécialisés.

(ii) pour assurer une approche multilingue le processus de traduction se divisera en trois étapes : analyse, transfert et génération.

La qualité linguistique d'une traduction est évidemment un des critères d'une importance majeure dans l'évaluation d'un système de TAO. Toutefois, les facteurs industriels en jeu laissent entendre que l'on doit aboutir à un compromis, entre le coût, l'efficacité et la qualité linguistique. Certains phénomènes linguistiques sont complexes, donc coûteux à traiter, alors qu'une solution technique, non-perfectionnée, moins exigeante, pourrait fournir, sur le plan commercial, un résultat adéquat.

Ainsi, les scientifiques favorisent une approche globale. Ils conçoivent un environnement facile à exploiter, offrant des outils divers à l'utilisateur (traducteur, rédacteur, etc.). Pour cela, ils prêtent une attention toute particulière aux outils de pré-édition et de post-édition. L'environnement définitif consistera en un poste de travail de

traducteur facile à utiliser pour la pré-édition et la post-édition.

B. Seite souligne certains critères primordiaux :

(i) Des structures ou des mots non-prévus ne devraient pas bloquer le processus de traduction. Le système devrait toujours fournir au moins une traduction possible.

(ii) Lorsque plusieurs traductions sont fournies, elles devraient être présentées d'une manière facile à comprendre pour l'utilisateur.

(iii) Afin de garantir que le système soit réutilisable, les chercheurs prévoient l'ajout de nouveaux éléments et l'amélioration d'anciens éléments.[123]

Étant donné que le système est conçu pour un domaine industriel, à l'aide d'EUROLANG, les scientifiques espèrent livrer non seulement un système efficace de TAO pour l'Europe mais aussi une plate-forme destinée aux diverses applications de TALN à une grande échelle.

---

[123] Seite et al, 1292.

6.2 TAO DU RÉDACTEUR

Pendant la période de 1993 à 1995 les Grenoblois comptent se diriger vers la TAO du rédacteur. Ils envisagent leur développement de la maquette LIDIA et pensent arriver à un démonstrateur de cette base en 1995. Leurs travaux sur la désambiguïsation au moyen de dialogue devraient mener à la définition d'un outil plus générique permettant aux linguistes de formuler eux-mêmes leurs propres stratégies de désambiguïsation.

L'équipe prévoit aussi la continuation des travaux sur les bases lexicales multilingues dans le cadre du projet NADIA. Ils progressent dans la formalisation et l'implémentation de ces bases, et recherche des outils rendant possible le passage entre NADIA et les sources lexicales accessibles.

Au niveau industriel, Boitet suggère vivement que le GETA se place dans la dynamique créée par les Japonais, en établissant une industrie de la TAO en Europe. Au Japon, la vente de stations de TAO est faite par Toshiba, Fujitsu, NEC, et Sharp. La vente actuelle permet aux producteurs de systèmes de couvrir au moins leurs coûts de développement courant.[124] Rien n'empêchera les sociétés japonaises d'établir des filiales de TAO en Europe et de développer des

---

[124] Ch. Boitet. "Mission d'étude sur la TAO au Japon". Document interne, Grenoble, 20-27 juin 1992, 16.

systèmes pour le marché européen. La poursuite d'Eurolang est souhaitable étant donné que l'élaboration des dictionnaires et des logiciels performants entre les langues européennes fournira aux Grenoblois un avantage par rapport aux Japonais.

Enfin, l'équipe compte prolonger sa participation au projet Eurolang. Leur apport principal portera sur l'étude de nouvelles techniques de génie logiciel pour le génie linguiciel, ainsi que sur la conception et le prototypage de nouveaux langages spécialisés.

7.0 CONCLUSION

Nous avons retracé l'évolution de la traduction automatique à Grenoble. Nous nous sommes attachée surtout à exposer les moyens que les chercheurs ont mis en oeuvre et à examiner les résultats obtenus, afin de mettre en lumière les principaux problèmes soulevés.

Si les préparatifs d'autres pays ont stimulé l'élaboration théorique de la traduction à Grenoble, c'est la continuité qui est le facteur essentiel de succès pour la recherche et le développement. Cette expérience concrète a permis de modifier les objectifs et de préciser un certain nombre de problèmes, liés surtout à la maintenance, au développement et à la signification.

En trente ans la théorie et la pratique de la traduction automatique ont tellement évolué que les conclusions du rapport ALPAC, qui en 1966-67 condamnaient la traduction automatisée, doivent être reformulées aujourd'hui. A l'époque, les limitations du matériel des ordinateurs, considérées conjointement avec les insuffisances des théories compréhensives du langage allaient encombrer les premières tentatives de traduction automatique.

La poursuite de la linguistique en tant que science révèle la difficulté des problèmes linguistiques mais affirme la faisabilité d'un nouvelle génération de systèmes de TAO.

Il s'ensuit que les objectifs changent dans le sens d'une réduction des ambitions des pionniers. Dans la mesure où l'on ne vise plus une "traduction complètement automatique de haute qualité", l'équipe à Grenoble prévoit des facilités d'intervention humaine : soit pré-édition, soit post-édition, ou bien en équipant le poste d'un écran multifenêtres. Il ne s'agit pas de systèmes de traduction purement automatiques, mais de systèmes de "traduction assistée par ordinateur".

Les techniques exposées ici sont assez variées. Les approches théoriques se reflètent dans les modèles successifs conçus en fonction de l'objectif recherché par l'opération traduisante. Lors de l'éclosion des travaux du CETA, les chercheurs commencaient partout à se référer aux nouvelles théories linguistiques. Le CETA, s'intéressant à la traduction multilingue, considère la notion de langue intermédiaire. Le système de Mel'chuk et Jholkovskij, suivant cette notion et construit sur la base de sémantique logique, est repris dans les procédés des Grenoblois. La recherche des équivalences sémantiques (au lieu des correspondances) entre les langues particulières doit faciliter le processus traduisant, surtout dans un contexte multilingue.

Le CETA était parti d'une théorie fondée sur le passage par un "langage pivot" hybride. La transformation de chaque phrase source en une formule de langage pivot doit donner lieu à des phrases cibles qui préservent le sens sans traiter les

contraintes morphologiques et syntaxiques des langues naturelles. Le "pivot" du processus traduisant réintroduit le problème de la définition du sens et de son interprétation.

Dès 1956 Yngve affirme que le processus de la TAO nécessite une séparation des tâches. Les travaux du CETA se basent sur les trois phases successifs : analyse, transfert et génération. Leurs programmes se distinguent aussi par le principe de la modélisation par niveaux. L'analyse se fait au moyen de trois modèles successifs (analyse morphologique, analyse syntaxique de surface, transformation en langage pivot). Enfin, leur système s'appuie déjà sur le principe de la séparation entre les données linguistiques et les programmes. Cela permet de clarifier et de bien séparer les concepts qui interviennent dans la description des grammaires et dictionnaires de ceux qui interviennent dans l'algorithme. Ainsi, à la différence de ceux de la première génération, leurs programmes ne dépendent pas des langues visées.

En 1971, les résultats de la traduction du "veilleur" font ressortir les sources de difficultés rencontrées lors des travaux effectués par CETA : les traductions fournies ne sont que des périphrases ; si l'analyse ne fournit pas un résultat complet au niveau sémantique, seule une traduction mot-à-mot est possible ; et enfin, lors du traitement des ambiguités, l'unité qui se limite à la phrase rend difficile la recherche des antécédents.

D'ailleurs, le phénomène de l'ambiguïté est un problème de taille pour le système du CETA. Ce système fournit une gamme complète d'hypothèses, et des conditions de bonne formation éliminent les résultats mal formés. Il serait un avantage important d'introduire avec chaque grammaire une heuristique qui accélère le processus d'analyse en évitant au maximum les constructions inutiles. Vauquois suggère l'utilité de garder des traces de la structure du texte d'entrée pour faciliter la génération.

A partir de 1971, à la suite d'un examen des expériences menées au CETA, le "GETA" est à la recherche des systèmes qui répondent aux critiques de l'approche pivote. ARIANE-78, un générateur de systèmes de TAO, permet d'écrire des maquettes de "TAO" multilingues au moyen des LSPL. Les structures "multiniveaux" d'ARIANE permettent de traiter les ambiguïtés lors de leur apparition ; les résultats partiels obtenus dans un seul niveau servent à résoudre les problèmes rencontrés dans les autres. Au moyen des langues spécialisées et un environnement interactif, ARIANE-78 fournit un outil relativement simple. Son développement se prolonge dans le cadre du projet EUROTRA, d'ESOPE, et du Projet National de TAO. Ainsi, des développements considérables ont été apportés au système : plusieurs outils ont été ajoutés et l'unité de traduction dépasse la phrase.

Devant les nouvelles demandes de traduction multilingue

de documents techniques de taille moyenne le GETA est à la recherche d'une nouvelle approche. La TAO du veilleur des années soixante fournit rapidement des traductions "grossières". Celle qui est réservée au réviseur, comme Ariane, effectue des traductions "brutes", destinées à être révisées. Elle n'est envisageable que pour les textes homogènes extensifs. Pourtant, les théoristes, à l'instar des praticiens, soulignent que le traducteur humain ne saurait traduire un texte se rattachant à un domaine précis sans connaissances spécialisées. Alors, le GETA s'oriente vers la TAO du "rédacteur". Sous la maquette LIDIA, au moyen d'une phase de "dialogue", l'auteur monolingue non-spécialiste coopère avec le système pour désambiguïser son texte. C'est la grande diffusion des micro-ordinateurs et l'apparition de l'hypertexte qui permettent d'envisager l'idée de mettre ce système à la portée de ces rédacteurs monolingues.

Remarquons que c'est toujours à la traduction multilingue que vise le GETA. Sous LIDIA, une même analyse profonde pourra être utilisée pour la traduction éventuelle vers plusieurs langues. Deuxièmement, afin de standardiser son texte, l'utilisateur se sert de langages "contrôlés". Nous doutons de l'éventualité pour le rédacteur d'associer aux champs le type de fragments qu'ils sont susceptibles de contenir.

Le projet LIDIA nécessite des dictionnaires contenant la

représentation des différents sens possibles des mots. Avec NADIA, afin de réduire le coût de rassemblement de bases lexicales, le GETA effectue un nouveau pas dans le développement de dictionnaires réutilisables. La base lexicale NADIA devra servir aux différentes applications de TALN et à l'utilisateur non-spécialiste.

A partir de 1980, dans le cadre du projet EUROLANG et pour la maquette LIDIA, l'équipe se limite surtout à la maintenance d'ARIANE-G5. Il devient évident qu'il n'est pas suffisant de savoir élaborer des systèmes de TAO, il faut savoir les utiliser. Pour préparer le "savoir-utiliser", les Grenoblois mettent en oeuvre des outils de TAO déjà disponibles et développent les outils supplémentaires indispensables, dans le cadre d'utilisations pilotes. Ils élaborent de nouveaux langages spécialisés et de nouvelles techniques de génie logiciel pour le génie linguiciel. Ils continuent aussi le développement de leurs bases lexicales multilingues (NADIA). La poursuite du projet LIDIA permettra éventuellement aux linguistes de formuler leurs propres stratégies de désambiguïsation.

Enfin, pour couvrir au moins leurs coûts de développement courants, Boitet suggère la création d'une véritable industrie en France. Les systèmes de traduction actuels élaborés à Grenoble exigent un utilisateur qui soit en liaison constante avec l'équipe de révision et développement. Il serait

souhaitable de concevoir des systèmes que l'utilisateur peut maîtriser et faire évoluer lui-même.

Cette nouvelle technologie exige beaucoup de temps et implique le travail non seulement des informaticiens, mais spécialement des linguistes et des lexicographes. Le GETA échange des données lexicales importantes préparées dans d'autres centres contre des outils informatiques qu'ils ont développés et procèdent à des échanges réguliers avec des équipes à Paris, à Montréal, à Pise, à Sarrebruck et à Moscou. Il s'agit d'équipes multidisciplinaires qui s'attachent à la poursuite non pas forcément d'un objectif commun mais des objectifs associés. De cette façon, le GETA serait en mesure d'acquérir des dictionnaires exceptionnellement riches en échange de leurs logiciels.

Plus les imperfections des solutions empiriques apparaissent, plus il devient clair que la traduction automatique devra s'appuyer sur des recherches mettant en jeu de vastes connaissances linguistiques. Le travail du GETA va de front avec une série de recherches sur la langue. Toutefois, la linguistique n'arrive pas à fournir de réponse positive globale au problème de la traduction automatique.

L'ambiguïté se révèle l'un des obstacles les plus difficiles à surmonter. L'exigence de fidélité de la traduction oblige le traducteur, ou bien la machine à traduire, à suivre de très près le texte. C'est ainsi que

l'ambiguïté de la langue se révèle pleinement. Si beaucoup d'ambiguïtés sont résolues au cours de la traduction automatique, il reste bien des cas irréductibles dans le choix d'une unité lexicale, d'une construction et dans les références pronominales.

La traduction automatique sensibilise donc à la nature intrinsèquement ambiguë de la langue. Le processus exige la spécification du sens des mots et ainsi la précision de leur contexte : un processus de désambiguïsation. En général, pour déterminer le sens d'un mot, la machine procède d'une série de recherches sur les mots environnants et sur la structure de la phrase. Le contexte dans les limites de la phrase est souvent considéré comme la limite du contexte susceptible d'être exploré ; à la fois pour des raisons technologiques et linguistiques. Jusqu'ici, même les machines les plus avancées ont besoin de données de la part des traducteurs humains pour résoudre les problèmes d'ambiguïté. Comme sous LIDIA, les connaissances "extra-linguistiques" du traducteur humain apportent des éléments d'information qui permettent de résoudre des ambiguïtés.

De même que l'on ne devrait concevoir une technique de traduction qui ne tienne pas compte de la syntaxe, l'on ne pourrait envisager une technique qui laisse peu de place à la sémantique, y compris le problème d'ambiguïté. Le CETA propose donc un langage pivot, des traits sémantiques et des

étiquettes. Le GETA offre des structures multiniveaux, des dictionnaires spécialisés, la pré-édition et la post-édition, des dictionnaires terminologiques et enfin une phase de désambiguïsation avec le rédacteur.

En examinant l'historique des essais tentés pour effectuer mécaniquement des traductions, nous discernons une nette évolution des problèmes. Qu'il n'y ait pas de solution universelle aux difficultés de la traduction automatique la présente expérience en témoigne clairement. La multiplicité des approches qui caractérisent les projets examinés ci-dessus souligne la complexité du matériau sur lequel on travaille. Le point le plus important est qu'il ne s'agit pas d'un outil de traduction automatique, mais d'un outil au sein du processus de traduction. Il ne s'agit plus de savoir si la TAO est réalisable, mais à l'aide de quelles techniques et dans quels domaines elle pourrait être plus efficace. Celui qui élabore un système de traduction doit invariablement faire un compromis entre la qualité du résultat et le coût de l'analyse.

La persévérance des Grenoblois a ouvert des optiques valables qui se révèlent d'un grand profit pour la linguistique. Les idées présentées ont joué un rôle capital dans l'évolution et dans l'ouverture des voies de recherche qui sont à peine explorées. La traduction automatique, aussi bien sur le plan des recherches que sur celui des réalisations

pratiques, loin d'être un sujet périmé, offre pour l'avenir une activité riche en nouvelles découvertes et en nouvelles applications.

OUVRAGES CONSULTÉS

ALPAC Language and Machines : Report of the Automated Language Processing Advisory Committee. Washington DC: National Academy of Sciences National Research Council, 1966.

Bachut, Daniel et Nelson Verastegui. "Software Tools for the Environment of a Computer Aided Translation System." Tenth International Conference on Computational Linguistics : Proceedings of COLING 84. Morristown (N.J.): Assoc. for Computational Linguistics, 1984.

--- et al. "Industrialisation d'un système de TAO français-anglais pour la documentation technique." Génie Linguistique 91, Versailles, 1991.

Baille, A. "Les systèmes logiques et les modèles sémantiques de langues naturelles." T.A. Informations. 1 (1968): 5-7.

--- et J. Rouault. "Un essai de formalisation de la sémantique des langues naturelles." T.A. Informations. 1 (1967): 1-7.

Bar-Hillel, Yehoshua. "A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation." Advances in Computers. 1 (1960): 158-163.

Blois, J. et al. Problèmes de la traduction automatique. Paris: Alabama UP, 1968.

Bloomfield, Leonard. Language. New York: Holt, 1933.

Boden, Margaret A. "The Meeting of Man and Machine." The Design of Informational Systems for Human Beings. Ed. Kevin P. Jones et Heather Taylor. London: Aslib Proceedings, 1981, 4-15.

Boitet, Christian. "La TAO à Grenoble en 1990 : 1980-1990 : TAO du réviseur et TAO du traducteur." La TAO à Grenoble en 1990. Grenoble: IMAG, 1990, 1-21.

---. "La TAO à Grenoble en 1990 : 1989- : vers la TAO du rédacteur (TAO personnelle)." La TAO à Grenoble en 1990. Grenoble: IMAG, 1990, 1-12.

---. "La TAO à Grenoble en 1990 : Présentation générale." La TAO à Grenoble en 1990. Grenoble: IMAG, 1990, 1-20.

---. Le logiciel ARIANE 78.5 du G.E.T.A. : Principes généraux, applications actuelles et futures. 5e Congrès national sur l'information et la documentation. 8-10 juin 1983. Grenoble: Information, Documentation, Transfert des Connaissances, 1983.

---. "Méthodes sémantiques en traduction automatique." T.A. Informations. 1 (1976): 3-42.

---. "Mission d'étude sur la TAO au Japon." Document interne, Grenoble, 20-27 juin 1992.

---. Motivations, Aims and Architecture of the LIDIA Project. MT Summit II : Machine Translation. 16-18 août, 1989, Munich, 50-54.

---. MT at Grenoble in 1990 : General Presentation. Proc. of ROCLING III : R.O.C. Computational Linguistics Conference III. Hsinchu: Tsing Hua University, 1990.

---. Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation. New Directions in Machine Translation. BSO congress, Budapest, août, 1988.

---. "Software and Lingware Engineering in Modern M(A)T Systems." Computational Linguistics : An International Handbook on Computer Oriented Language Research and Applications. Berlin: Walter de Gruyter, 1989, 670-682.

---. "TA et TAO à Grenoble... 32 ans déjà!" T.A.L. (revue semestrielle de l'ATALA). 33:1-2, Spécial Trentenaire (1992): 45-84.

---. "The French National MT-Project : Technical Organization and Translation Results of CALLIOPE-AERO." Computers and Translation. 1:4 (1985): 239-267.

---. Towards Personal MT : General Design, Dialogue Structure Potential Role of Speech, Text Encoding. ROCLING-III. Taipeh, 1990, 63-70.

---. "Twelve Problems for Machine Translation." International Conference on Current Issues in Computational Linguistics, Penang, 12-14 juin 1991.

--- et H. Blanchon. "La TA fondée sur le dialogue pour auteurs monolingues et le projet LIDIA." Rapport de recherche. GETA, IMAG, UJF et CNRS, Grenoble, juin 6, 1993.

--- et René Gerber. Expert Systems and Other New Techniques in MT Systems. International Conference on Computational Linguistics : Proceedings of COLING 84. Morristown, N.J.: Assoc. for Computational Linguistics, 1984.

--- et N. Nedobejkine. "L'informatique au service de la linguistique : Illustration sur le développement d'un atelier de traduction automatisée." La recherche française par ordinateur en langue et littérature. Actes du colloque, l'Université de Metz, juin 1983. Genève: Slatkine, 1985.

--- et F. X. Tchéou. On a Phonetic and Structural Encoding of Chinese characters in Chinese texts. ROCLING-III. Taipeh, 1990, 73-80.

---, Pierre Guillaume, et Maurice Quézel-Ambrunaz. "Ariane-78, an integrated environment for automated translation and human revision." Proc. COLING-82, North-Holland, Ling. series 47, Prague, 1982, 19-27.

---, Pierre Guillaume, et Maurice Quézel-Ambrunaz. "Implementation and Conversational Environment of ARIANE 78.4, an Integrated System for Automated Translation and Human Revision." Sprachen und Computer: Festschrift zum 75. Dudweiler: AQ-Verlag, 1982, 225-236.

--- et al. Le point sur la participation du GETA à EUROTRA. Document interne, Grenoble, 1988, 2.

Cary, E. "Mécanismes et traduction." Babel. 2:3 (1956): 102-107.

Ceccato, Silvio. "Operational Linguistics and Translation." Linguistic Analysis and Programming for Mechanical Translation. Milan: University of Milan Italy, Giangiacomo Feltrinelli Editore, 1960, 11-80.

Chauché, J. "Les Systèmes A.T.E.F. et C.E.T.A." T.A.Informations. 2 (1975): 27-38.

---. Transducteurs et arborescences. Étude et réalisation de systèmes appliqués aux grammaires transformationnelles. Thèse d'État, Grenoble, 1974.

Chomsky, N. Syntactic Structures. The Hague: Mouton, 1957.

Colmerauer, A. "Les systèmes-Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur." Projet TAUM, Université de Montréal, janvier 1971.

Coughlin, Josette M. "Artificial Intelligence and Machine Translation : Present Developments and Future Prospects." Babel. 34:1 (1988): 3-9.

Delavenay, Emile. La machine à traduire. Paris: PUF, 1972.

---. "Quelques réflexions après dix ans..." T.A. Informations. 2 (1969): 62-64.

De Roeck, A. "Linguistic Theory and Early Machine Translation." Machine Translation Today: The State of the Art. Ed. Margaret King. Edinburgh: Edinburgh UP, 1987, 38-57.

Detemple, Albert. "La traduction assistée par ordinateur (TAO) au Centre de Documentation Scientifique et Technique (CDST) du Centre National de la Recherche Scientifique (CNRS)." Multilingua. 2:3 (1983): 189-194.

Galisson, R. et D. Coste. Dictionnaire de didactique des langues. Paris: Hachette, 1976.

Garvin, Paul L. On Machine Translation: Selected Papers. The Hague: Mouton, 1972.

---. ed. Natural Language and the Computer. New York: McGraw-Hill, 1963.

Gougenheim, G. Problèmes de la traduction automatique. Alabama: Alabama UP, 1968.

Guilbaud, J.-Ph. Descripteur Linguistique multiniveau et génération de texte en Ariane-78. Présenté pour la Journée ATHENA sur la traduction automatique. Université de Liège, 5 mai, 1987.

--- et N. Nedobejkine. Rapport sur Ariane-G5 : Point de vue utilisateur linguiste. Document interne. EUROLANG, Grenoble, 27-28 février 1992, 12-16.

Guillemin-Flescher, Jacqueline. Syntaxe comparée du français et de l'anglais : problèmes de traduction. Paris: Éditions Ophrys, 1981.

Hutchins, W.J. Machine Translation: Past, Present, Future. New York: Halsted Press, 1986.

---. "The Evolution of Machine Translation Systems." Practical Experience of Machine Translation : Proceedings of a Conference. Ed. Veronica Lawson. Amsterdam: North-Holland, 1982.

Johnson, Rod. "Contemporary Perspectives in Machine Translation." Contrastes : Revue de l'Association pour le développement des études contrastives. A4 (1984): 141- 155.

Josselson, H.H. et al. Fourteenth (Final) Annual Report on Research in Computer-Aided Translation Russian-English, Wayne State University, avril 1972.

Kelly, Ian D.K., ed. Progress in Machine Translation: Natural Language and Personal Computers. UK: Sigma Press, 1989.

Kingscott, Geoffrey. "Buys B'Vital : Relaunch of French National MT Project." Language International. 2:2 (1990): 3-6.

Kittredge, R. Analyzing Language in Restricted Domains : Sublanguage Description and Processing. Grishman R. & Kittredge R., eds. New-Jersey: Lawrence Erlbaum, 1986.

---. "Sublangage - Specific Computer Aids to Translation : A Survey of the Most Promising Application Areas." Contract no 2-5273, Université de Montréal et Bureau des Traductions, mars 1983.

--- et John Lehrberger, eds. Sublanguage: Studies of Language in Restricted Semantic Domain. Berlin: Walter de Gruyter, 1982.

Kulas, Jack, James H. Fetzer et Terry C. Rankin, eds. Philosophy, Language, and Artificial Intelligence. Dordrecht: Kluwer Academic, 1988.

Ladmiral, J.-R. Traduire : théorèmes pour la traduction. Paris: Payot, 1979.

Lafourcade, M. "ODILE-2, un outil pour traducteurs occasionnels sur Macintosh." Colloque "L'environnement traductionnel : La station de travail du traducteur de l'an 2001.", UREF & AUPELF, Mons, avril 1991, PU du Québec, 95-108.

Lamb, Sydney M. Outline of Stratificational Grammar. Washington, D.C.: Georgetown UP, 1966.

--- et W. Jacobson. "A High-speed Large Capacity Dictionary System." Mechanical Translation, vol 6, novembre 1961.

Larose, Robert. Théories contemporaines de la traduction. 2e éd. Québec: Québec UP, 1989.

Lawson, Veronica ed. Parsing Natural Language. New York: Academic Press, 1983.

Locke, William et al. Machine Translation of Languages. New York: MIT UP et Wiley, 1955.

Loffler-Laurian, Anne-Marie. "Pour une typologie des erreurs dans la traduction automatique." Multilingua. 2:2 (1983): 65-78.

---. "Traduction automatique et périphériques : Évaluation, post-édition, attitudes, formation." Contrastes : Revue de l'Association pour le développement des études contrastives. A4 (1984): 43-67.

Ludskanov, A. "A propos des problèmes théoriques de la traduction." T.A. Informations. 2 (1967): 106-112.

Macdonald, R.R. Georgetown University Machine Translation Project. General Report 1952-1963. Washington, D.C., 1963.

Maegaard, Bente. "Eurotra : The Machine Translation Project of the European Communities." Literary and Linguistic Computing. 3:2 (1988): 61-65.

--- et Sergei Perschke. "Eurotra : General System Design." Machine Translation. 6:2 (1991): 73-82.

Margot, Odette. "Le second congrès de linguistique appliquée." T.A.Informations. 1 (1970): 24-32.

Mel'cuk, Igor A. "Meaning-Text Models : A Recent Trend in Soviet Linguistics." Annual Review of Anthropology. 10 (1981): 27-62.

--- et A.K. Zholkovskij. "Towards a Functioning 'Meaning-Text' Model of Language." Linguistics : An International Review. 57 (1970): 10-47.

Meyer, I., B. Onyshkevych et L. Carlson. "Lexicographic Principles and Design for Knowledge-Based Machine Translation", Carnegie Mellon University, Technical Report no CMU-CMT-90-118, 13 août 1990.

Mounin, Georges. La machine à traduire : histoire des problèmes linguistiques. The Hague: Mouton, 1964.

---. Les problèmes théoriques de la traduction. Paris: Éditions Gallimard, 1976.

---. Linguistique et traduction. Bruxelles: Dessart et Mardaga, 1976.

Moyne, John A. Understanding Language: Man or Machine. New York: Plenum Press, 1985.

Nagao, Makoto. "Future Directions of Machine Translation." Prague Bulletin of Mathematical Linguistics. 51 (1989): 20-24.

---. "La traduction automatique." La Recherche. 150 (1983): 1530-1541.

---. Machine Translation: How Far Can It Go? Trans. Norman D. Cook. Oxford: Oxford UP, 1989.

Nida, Eugene A. Towards a Science of Translating. Leyde: Brill, 1964.

Nirenburg, Sergei, éd. Machine Translation: Theoretical and Methodological Issues. Cambridge (UK): Cambridge UP, 1987.

Panov, D., A. Ljapunov et I. Moukhine. La traduction automatique. 7 (1958): 162-193.

Peccoud F. "The Aims of the French National Project of Computer-Aided Translation." In International Forum on Information and Documentation. 13:1 (1988): 11-13.

Pigott, I. "Systran : A Key to Overcoming Language Barriers in Europe." Multilingua. 2:3 (1983): 149-156.

---. "The Importance of Feedback from Translators in the Development of High-Quality Machine Translation." Practical Experience of Machine Translation: Proceedings of a Conference. Éd. Veronica Lawson. Amsterdam: North-Holland, 1982, 61-65.

Richardson, S.D. Enhanced Text Critiquing using a Natural Language Parser: the CRITIQUE system. IBM Research Report RC 11332, Thomas J. Watson Research Center, Yorktown Heights, 1985.

Roberts, A.H. et M. Zarechnak. "Mechanical Translation." Current Trends in Linguistics, vol. 12: Linguistics and Adjacent Arts and Sciences, pt.4. The Hague: Mouton, 1974, 2825-2868.

Sampson, C. "MT: A Nonconformist's View of the Art." Machine Translation Today: The State of the Art. Ed. Margaret King. Edinburgh: Edinburgh UP, 1987, 91-110.

Schneider, T. "The Metal System." MT Summit III, Washington, 1991.

Seite, B. et al. Presentation of the Eurolang Project. Proc. of COLING-92. Nantes, 23-28 août, 1992.

Sérasset, G. NADIA - Une nouvelle approche: les dictionnaires interlingues par acceptions. Rapport de DEA d'informatique, UJF et INPG, GETA, IMAG, juin 1991.

--- et E. Blanc. Une approche fondée sur les acceptions pour les bases lexicales multilingues. Proc. IIIèmes journées scientifiques "Traductique-TA-TAO", Montréal, octobre 1993, Presses de l'Université de Montréal.

Shann, P. "Machine Translation: A Problem of Linguistic Engineering or of Cognitive Modelling?" Machine Translation Today: The State of the Art. Ed. Margaret King. Edinburgh: Edinburgh UP, 1987, 71-90.

Slocum, Jonathan. "A Survey of Machine Translation : Its History, Current Status, and Future Prospects." Machine Translation Systems. Éd. Jonathan Slocum. Cambridge (Mass): Cambridge UP, 1988.

Somers, H.L. "Eurotra Special Issue." Multilingua. 5:3 (1986): 125-127.

Steiner, George. After Babel: Aspects of Language and Translation. London: Oxford UP, 1975.

TEI : Guidelines for the encoding and the interchange of machine-readable texts. ACH-ACL-ATC., juillet 1990.

Tesnière, Lucien. Éléments de Syntaxe Structurale. Paris: Klincksieck, 1969.

Tomita, Masaru. Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems. Boston: Kluwer Academic, 1986.

Tsutsumi, T et al. "Example-Based Approach to Machine Translation." Proc. Premières journées franco-japonaises sur la traduction assistée par ordinateur. Ambassade de France au Japon, Tokyo, Japon. 15-16 mars 1993, 1:1, 1993, 161-169.

Van Slype, Georges. "Evaluation du système de traduction automatique SYSTRAN anglais-français, version 1978, de la Commission des communautés européennes." Babel. 3:25 (1979): 57-162.

Vauquois, Bernard. Bernard Vauquois et la TAO, vingt-cinq ans de Traduction Automatique : ANALECTES. Éd. Boitet Ch. Grenoble, Champollion & GETA, 1988.

---. "Dix ans d'ATALA : De la traduction automatique au traitement automatique des langues." T.A. Informations. 2 (1969): 57-61.

---. La traduction automatique à Grenoble. Les Documents de Linguistique Quantitative 24. Paris: Dunod, 1975.

---. "Présentation du Centre d'Études pour la Traduction Automatique (C.E.T.A.) du Centre National de la Recherche Scientifique." T.A. Informations. 1 (1966): 1-18.

---. "Structures profondes et traduction automatique : Le système du C.E.T.A." Revue Roumaine de Linguistique. Tome XIII, No.2. (1968): 105-129.

---. Syntaxe et interprétation. Proceedings of the 1965 International Conference on Computational Linguistics, New York, mai 1965.

---. "Traduction assistée par ordinateur. Formation de spécialistes : Préparation du transfert technologique." Projet ESOPE, Contrat ADI, mai 1981.

--- et S. Chappuy. "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, August 14-16, 1985.

--- et al. "Définition d'une méthode de travail d'équipe linguistique." Projet ESOPE, Contrat ADI, novembre 1982.

--- et al. "Une notation des textes hors des contraintes morphologiques et syntaxiques de l'expression." T.A.Informations. 1 (1970): 1-20.

Veillon, G. Consultation d'un dictionnaire et analyse morphologique en Traduction Automatique. Thèse de 3ème cycle, Université de Grenoble I, juin 1962.

---. "Description du langage pivot du système de traduction automatique du C.E.T.A." T.A. Informations. 1 (1968): 8-17.

Véronis, J., N.M. Ide et S. Harié. Construction automatique de grands réseaux de neurones pour la désambiguïsation du langage naturel, 10èmes journées Systèmes Experts et leur application. AVIGNON'90, Avignon, 28 mai-1 juin 1990, 105-117.

Vinay, J.-P. et J. Darbelnet. Stylistique comparée du français et de l'anglais: Méthode de traduction. Paris: Didier, 1958.

Walker, P.A. "A Commission of the European Communities User Looks at Machine Translation." Tools for the Trade: Translating and the Computer 5. Éd. Veronica Lawson. London: Aslib, 1985, 145-163.

Warwick, S. "An Overview of Post-ALPAC Developments." Machine Translation Today: The State of the Art. Éd. Margaret King. Edinburgh: Edinburgh UP, 1987, 22-37.

Weaver, W. "Translation." Éds. W.N.Locke, et A.D.Booth. Machine Translation of Languages. New York: Wiley, 1955, 15-23.

Wehrli, E. "Recent Developments in Theoretical Linguistics and Implications for Machine Translation." Machine Translation Today: The State of the Art. Éd. Margaret King. Edinburgh: Edinburgh UP, 1987, 58-70.

Winograd T. "Procedural Model of Language Understanding". Computer Models of Thought and Language. Éds. Shank & Colby, Freeman, San Francisco, 1973. 152-186.

---. "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language." AI-TR-17, MIT, Cambridge, Mass., jan. 71.

Woods, William. Transition Network Grammars for Natural Language Analysis. CACM 13:10, oct.1970, 591-606.

Yngve, Victor. "A Framework for Syntactic Translation." Mechanical Translation 4:3 (1957): 59-65.

---. "Implications of Mechanical Translation Research." Proceedings of the American Philosophical Society. 108:4 (1964): 275-281.

---. "Report." Proceedings of the VIII International Congress of Linguists. Oslo: Oslo UP, 1958, 510-518.

Zarechnak, M. "The History of Machine Translation." Henisz-Dostert, B. et al. Machine Translation. The Hague: Mouton, 1979, 3-87.

Zholkovskij, I.A. et I.A. Mel'chuk. "Sur la synthèse sémantique." T.A. Informations. 2 (1970): 1-85.

DÉFINITIONS

Nos définitions restent opératoires dans le cadre de nos analyses. Il est rare en effet qu'un phénomène puisse être défini totalement par une catégorie. Les termes utilisés ici se chevauchent et se contredisent souvent parmi les chercheurs. En général, nous adoptons les définitions du Groupe d'Étude pour la Traduction Automatique.

Tout comme le GETA, nous distinguons les "maquettes", qui sont des programmes développés pour l'expérimentation réduite de techniques ou d'architectures nouvelles (comme LIDIA) ; les "prototypes", qui sont des expériences de laboratoire portant sur des corpus, des grammaires ou des dictionnaires relativement limités, par exemple 5 à 10 000 termes, et n'ayant pas été testés dans des conditions opérationnelles, i.e. sur un flux de textes nouveaux (comme EUROTRA) ; et les "systèmes", qui sont des applications opérationnelles, ou de générateurs supportant de telles applications, avec (sauf pour le cas particulier de METEO) au moins 20 à 30 000 termes, 300 à 400 pages pour les grammaires d'analyse, et application à des flux de textes réels (comme ARIANE, LOGOS, METEO, METAL, SUSY, SYSTRAN).

A la fin des années soixante-dix, le GETA introduit le terme "Traduction Assistée par Ordinateur" (TAO). Nous

adoptons cette formule pour parler des possibilités élargies d'informatisation de la traduction, impliquant la TA du veilleur, la TA du réviseur, la TA interactive (projet LIDIA), et enfin la TA(O) du traducteur (ou THAO Traduction Humaine Assistée par Ordinateur, dans laquelle on fournit un "poste de travail" adapté au traducteur-réviseur).

Dans l'analyse du niveau de l'intervention humaine, nous employons les divisions conçues par Masaru Tomita[125] :

(i) Rejet - Les phrases que le système n'arrive pas à traiter sont rejetées. Elles sont ensuite traduites par un traducteur humain.

(ii) Pré-édition - Les textes de départ sont modifiés par les humains. De cette façon, les textes ne se composent que de la syntaxe et du vocabulaire que le système peut traiter. Selon Vauquois, la pré-édition ne consiste pas à transformer le texte pour le mettre sous une forme acceptable par le système automatique : "une annotation facultative et pas nécessairement complète du texte source au moyen d'un certain nombre de marqueurs, afin d'aider efficacement le système automatique à résoudre

---

[125] Masaru Tomita. Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems. Boston: Kluwer Academic, 1986.

les ambiguités les plus difficiles".[126]

(iii) Post-édition - La méthode la plus utilisée. Le système "accepte" les textes de départ non édités et fournit aux réviseurs la sortie sur imprimante du texte traduit par le processus automatique.

(iv) Interactive - Le système ne nécessite ni pré-édition, ni post-édition ; au contraire, le secours humain interactif a lieu au cours du processus de traduction.

---

[126] Vauquois. ANALECTES, 515.

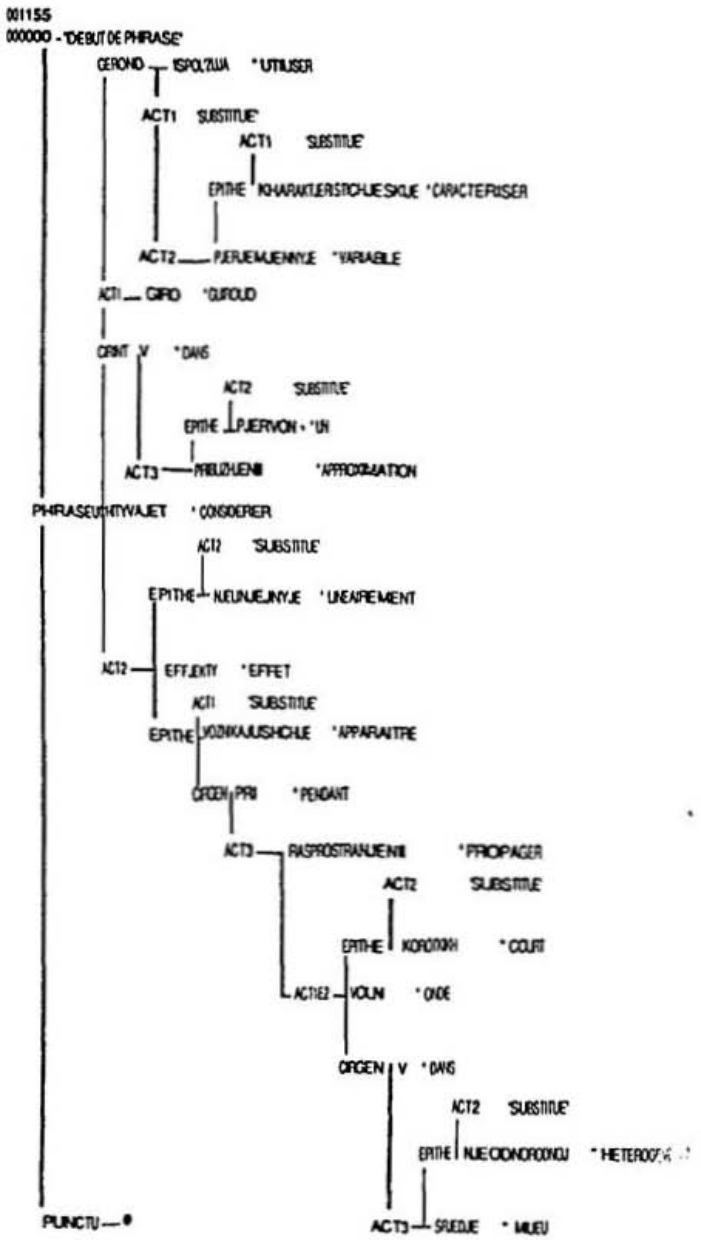
ANNEXES

Annexe 1. Un fragment du texte russe traduit automatiquement sur l'ordinateur IBM 7044.

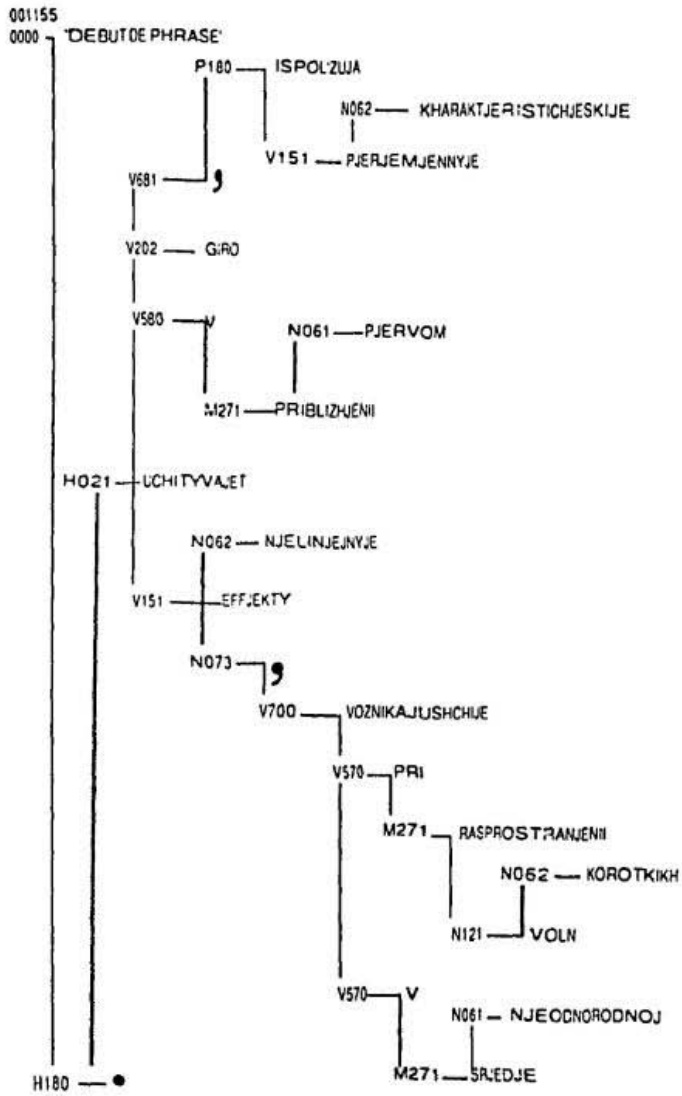
SII/СПОЛЬЗУЯ ХАРАКТЕРИСТИЧЕСКИЕ ПЕРЕМЕННЫЕ, SГ/ИРО
В ПЕРВОМ ПРИБЛИЖЕНИИ УЧИТЫВАЕТ НЕЛИНЕЙНЫЕ
ЭФФЕКТЫ, ВОЗНИКАЮЩИЕ ПРИ РАСПРОСТРАНЕНИИ
КОРОТКИХ ВОЛН В НЕОДНОРОДНОЙ СРЕДЕ.

La zone soulignée correspond à la phrase dont les pages suivantes illustrent les résultats à la sortie des principaux modèles.

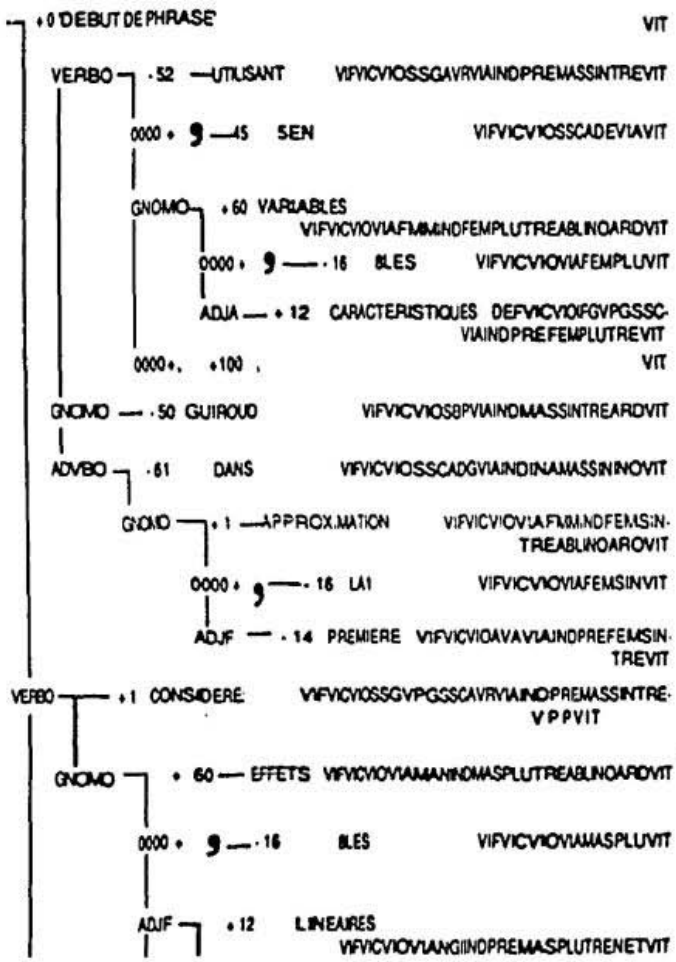
Annexe 2. La structure de dépendances sous forme d'une arborescence fournie par l'analyse syntaxique.



Annexe 3. La représentation "pivote" produite par l'analyse profonde avec les équivalents français.



Annexe 4. La construction syntaxique française de surface.



Annexe 5. La phrase soulignée est le résultat de traduction brute de la phrase de l'annexe 1.

... PENDANT CETTE HYPOTHESE, LES PERTURBATIONS, FAIBLES DANS LE MILIEU HETEROGENE SE PROPAGENT DE MANIERE QUASI-UNIDIMENSIONNELLE LE LONG DES RAYONS CARACTERISTIQUES. DANS LES TRAVAUX DE RYJOV, QUI A EXAMINE LA PROPAGATION DE L'ONDE AVEC UN PROFIL TRIANGULAIRE DE LA PRESSION EXCEDANTE, LA METHODE DE LANDAU A ETE GENERALISEE AU CAS DU MILIEU HETEROGENE ET LA LOI ASYMPTOTIQUE DE L'AMORTISSEMENT D'ONDES NON STATIONNAIRES ET STATIONNAIRES DE CHOC DANS LES MILIEUX HETEROGENES A ETE OBTENUE. RYJOV A REMARQUE QUE, PENDANT L'ETUDE DE L'AMORTISSEMENT D'ONDES DE CHOC, IL FAUT CONSIDERER LA DIVERGENCE, LIEE A SON HETEROGENEITE, QUI SE PRODUIT DEJA DANS L'APPROXIMATION DE L'ACOUSTIQUE GEOMETRIQUE, DES SURFACES CARACTERISTIQUES DU MILIEU NON PERTURBE. DANS CES TRAVAUX CEPENDANT SEULE LA LOI ASYMPTOTIQUE DE L'AMORTISSEMENT D'ONDE DE CHOC DANS LE MILIEU HETEROGENE SUR LES GRANDES DISTANCES DE LA SOURCE DE SON APPARITION A ETE OBTENUE, ET LA DEPENDANCE DE SON AMPLITUDE PAR RAPPORT A LA FORME DU CORPS, OU PAR RAPPORT A LA DISTRIBUTION INITIALE DE LA PRESSION EXCEDANTE N'A PAS ETE ETABLIE. TELLE LIAISON A ETE ETABLIE DANS LE TRAVAIL COLLECTIF DE FRIEDMAN, DE KANE ET SIGALLA (12) ET DANS LE TRAVAIL DE GUIROUD (13). DANS LE TRAVAIL (12), LA PROPAGATION QUASI-UNIDIMENSIONNELLE DES PERTURBATIONS DANS LE TUBE RADIAL EST EXAMINEE. LA SOLUTION OBTENUE DANS ELLE DEPEND D'UNE CONSTANTE INDETERMINEE QUELCONQUE, QUI SE TROUVE DE LA CONDITION DE LA JONCTION DE SOLUTION OBTENUE DANS LE CAS DU MILIEU HOMOGENE AVEC UNE FORMULE ASYMPTOTIQUE DE WHITHAM POUR L'ONDE EN FORME DE
. GRACE A CELA DANS LA FORMULE ASYMPTOTIQUE POUR L'AMPLITUDE DE L'ONDE DE CHOC, LA DEPENDANCE PAR RAPPORT A LA FORME DU CORPUS APPARAIT. DANS LE TRAVAIL DE GUIROUD, POUR L'ETUDE DE L'AMORTISSEMENT DE PERTURBATIONS, PROVOQUEES PAR LE MOUVEMENT DU CORPS, DANS LE MILIEU HETEROGENE, LA METHODE DES DEVELOPPEMENTS INTERNES ET EXTERNES EST APPLIQUEE. DANS LA PREMIERE APPROXIMATION EN UTILISANT LES VARIABLES CARACTERISTIQUES GUIROUD CONSIDERE LES EFFETS NON LINEAIRES, QUI APPARAISSENT PENDANT LA PROPAGATION DES ONDES COURTES DANS LE MILIEU HETEROGENE. DANS CE TRAVAIL LE CAS ASYMPTOTIQUE DE L'AMORTISSEMENT D'ONDES DE CHOC OU SUR LES GRANDES DISTANCES DE LA SOURCE DES PERTURBATIONS L'ONDE EN FORME DE
SE FORME A ETE EXAMINEE. DANS LE TRAVAIL DE GUIROUD L'INFLUENCE DE LA DIVERGENCE DES SURFACES CARACTERISTIQUES DU MILIEU NON PERTURBE SUR L'AMORTISSEMENT D'ONDES DE CHOC MENTIONNEE PLUS HAUT N'A PAS ETE CONSIDERE .