UNIVERSITÉS FRANCOPHONES

André Clas
Université de Montréal

Pierrette Bouillon
ISSCO, Université de Genève

TA-TAO:
RECHERCHES DE POINTE ET APPLICATIONS IMMEDIATES

Actes du Colloque de Montréal
1993

AUPELF·UREF

actualité scientifique

FMA

TA-TAO : RECHERCHES DE POINTE ET APPLICATIONS IMMÉDIATES

ISBN 2-909611-09-4

Bibliothèque nationale du Québec
Bibliothèque nationale du Canada
Bibliothèque nationale de France
Imprimé au Liban.

TA-TAO :
RECHERCHES DE POINTE ET APPLICATIONS IMMÉDIATES

Troisièmes Journées scientifiques du réseau thématique
«Lexicologie, Terminologie, Traduction»
Montréal, 30 septembre, 1er et 2 octobre 1993

Sous la direction de :

André CLAS, Université de Montréal
Pierrette BOUILLON, ISSCO Université de Genève

1994

FMA
Beyrouth

AUPELF • UREF
B.P 400, succ. Côte-des-Neiges
Montréal (Québec) Canada
H3S 2S5

Avant-propos

La diffusion de l'information scientifique et technique est un facteur essentiel du développement. Aussi dès 1988, l'Agence francophone pour l'enseignement supérieur et la recherche (AUPELF-UREF), mandatée par les Sommets francophones pour produire et diffuser revues et livres scientifiques, a créé la collection Universités francophones.

Lieu d'expression de la communauté scientifique de langue française, Universités francophones vise à instaurer une collaboration entre enseignants et chercheurs francophones en publiant des ouvrages, coédités avec des éditeurs francophones, et largement diffusés dans les pays du Sud, grâce à une politique tarifaire préférentielle.

Quatre séries composent la collection :

– Les manuels : cette série didactique est le cœur de la collection. Elle s'adresse à un public de deuxième et troisième cycles universitaires et vise à constituer une bibliothèque de référence couvrant les principales disciplines enseignées à l'université.

– Sciences en marche : cette série se compose de monographies qui font la synthèse des travaux de recherche en cours.

– Actualité scientifique : dans cette série sont publiés les actes de colloques organisés par les réseaux thématiques de recherche de l'UREF.

– Prospectives francophones : s'inscrivent dans cette série des ouvrages de réflexion donnant l'éclairage de la francophonie sur les grandes questions contemporaines.

Notre collection, en proposant une approche plurielle et singulière de la science, adaptée aux réalités multiples de la francophonie, contribue efficacement à promouvoir la recherche dans l'espace francophone et le plurilinguisme dans la recherche internationale.

Professeur Michel Guillou
Directeur général de l'AUPELF
Recteur de l'UREF

Sommaire

Partie II

Liste des auteurs

Blanc, Étienne, GETA (UJF & CNRS), IMAG-campus, 150, rue de la Chimie, BP 53 X, F-38041 Grenoble Cedex, France

Blanchon, Élisabeth, Centre de terminologie et de néologie CNT, INaLF - CNRS, 27, rue Damesme, 75013 Paris, France

Boitet, Christian, GETA (UJF & CNRS), IMAG-campus, 150, rue de la Chimie, BP 53 X, F-38041 Grenoble Cedex, France

Bouillon, Pierrette, ISSCO, Université de Genève, 54 route des Acacias, CH-1227 Genève, Suisse

Burkhard, Patrick, Union de Banques Suisses, GRCT, Bahnhofstrasse 45, CH-8021 Zürich, Suisse

Danlos, Laurence, TALANA, Université Paris VII, case 7003, 2 Place Jussieu, F-75251, Paris Cedex 05, France

Dauphin, Éva, Groupe TA-TAO, AÉROSPATIALE, Centre commun de recherches Louis-Blériot, Département Informatique Documentation, 12, rue Pasteur, BP 76, 92152 Suresnes Cedex, France

Degand, Liesbeth, Université Catholique de Louvain, Faculté de psychologie - Unité EXCO, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgique

Durieux, Christine, Université Paris III - Sorbonne Nouvelle, ESIT, Centre universitaire Dauphine, 75116 Paris, France

Estival, Dominique, ISSCO, Université de Genève, 54, route des Acacias, CH-1227 Genève, Suisse

Farwell, David, Computing Research Laboratory, Université New Mexico State, Las Cruces, NM 88003, États-Unis

Gentilhomme, Yves, 111, Grande-rue, 25000 Besançon, France

Gross, Gaston, Université Paris-Nord, Laboratoire de Linguistique Informatique, avenue J.B. Clément, F-93430 Villetaneuse, France

Habaili, Hussein, BLOC D2, App. 101, Cité la Forêt, Radès 2040, Tunisie

Hutchins, John, The Library, Université d'East Anglia, Norwich, NR4 7TJ, Royaume-Uni

Iordanskaja, Lidija, Université de Montréal, Département de linguistique et de traduction, C.P. 6128, succursale Centre-ville, Montréal (Québec) H3C 3J7, Canada

L'Homme, Marie-Claude, Lexi-tech inc., 10, avenue Dawson, Moncton (Nouveau-Brunswick) E1A 6C3, Canada

Lafourcade, Mathieu, GETA (UJF & CNRS), IMAG-campus, 150, rue de la Chimie, BP 53 X, F-38041 Grenoble Cedex, France

Lazourski, Alexandre, Institute for Information Transmission Problems, Russian Academy of Sciences, 19, Ermolova street, GSP-4, Moscou 101447, Russie

Lejosne, Jean-Claude, Université de Metz, UFR Lettres - Campus technopôle, Metz 2000, 7, rue Marconi, 57070 Metz, France

Lerat, Pierre, LSHS, CELEX, Université Paris XIII, avenue J.B. Clément, F-93430 Villetaneuse, France

Lerot, Jacques, Groupe de recherche GENESE, Institut de linguistique UCL, place Blaise-Pascal 1, B-1348 Louvain-la-Neuve, Belgique

Lonsdale, Deryle, Center for Machine Translation, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15215-3890, États-Unis

McHaffie, Bruce, École de traduction et d'interprétation, Université d'Ottawa, 52, rue Université, Ottawa (Ontario) K1N 6N5, Canada

Meijer, Siety, Université d'Essex, CL/MT Group, Département de langue et de linguistique, Wivenhoe Park, Colchester CO4 3SQ, Angleterre

Mel'čuk, Igor, A, Université de Montréal, Département de linguistique et de traduction, C.P. 6128, succursale Centre-ville, Montréal (Québec) H3C 3J7, Canada

Melby, Alan, 1223 Aston Avenue, Provo, 84604 Utah, États-Unis

Meyer, Ingrid, École de traduction et d'interprétation, Université d'Ottawa, 52, rue Université, Ottawa (Ontario) K1N 6N5, Canada

Nirenburg, Sergei, Center for Machine Translation, Université Carnegie Mellon, 5000 Forbes Ave, Pittsburgh, PA 15215-3890, États-Unis

Quirion, Jean, Services gouvernementaux Canada, Services de traduction, Ottawa (Ontario) KlA 0M5, Canada

Rabenilaina, Roger-Bruno, 30, Cité des Professeurs, Fort-Duchesne, 101, Antananarivo, Antananarivo, Madagascar

Rinsche, Adriane, The Language Technology Centre, 22 Cranleigh Gardens, Kingston, Surrey, KT2 5TX, Grande-Bretagne

Russell, Graham, ISSCO, Université de Genève, 54 route des Acacias, CH-1227 Genève, Suisse

Sérasset, Gilles, GETA (UJF & CNRS), IMAG-campus, 150, rue de la Chimie, BP 53 X, F-38041 Grenoble Cedex, France

Tomokiyo, Mutsuko, ATR Interpreting Telecommunications Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japon

Van Campenhoudt, Marc, ISTI, TERMISTI, Département de linguistique française, 34, rue J. Hazard, 1180 Bruxelles, Belgique

Warwick-Armstrong, Susan, ISSCO, Université de Genève, 54 route des Acacias, CH-1227, Genève, Suisse

Wehrli, Éric, Département de linguistique générale et de linguistique française, Université de Genève, 1211 Genève 4, Suisse

Wilks, Yorick, Computing Research Laboratory, Université New Mexico State, Las Cruces, NM 88003, États-Unis

Membres du comité du réseau «LTT»

Préface

André Clas

Université de Montréal, Canada

Si l'on se fie au nombre de participants aux Troisièmes Journées scientifiques du réseau thématique Lexicologie, Terminologie, Traduction, on ne peut être que frappé par ce regain d'intérêt pour les questions de traduction par ordinateur. En effet, plus de deux cent soixante personnes ont participé assidûment à cette rencontre et plus de quarante personnes ont assisté au séminaire de mise à niveau qui précédait. Sans doute les intérêts de tous ces participants peuvent-ils être différents : les uns veulent chercher une solution à la quantité de traductions à effectuer, une solution nouvelle, rapide, efficace et peu coûteuse ; les autres se préoccupent plus particulièrement des recherches théoriques et pour faire face aux défis posés par la TA ou la TAO. Le thème des Journées était bien centré sur ces deux aspects – Recherches de pointe et Applications immédiates. C'est ainsi que l'on fait d'ailleurs progresser la connaissance et que l'on stimule la recherche : qui fait quoi en ce moment et comment le fait-il ?

On sait que la naissance de l'ordinateur et le désir de nouvelles utilisations sont à l'origine des premières recherches sur la traduction par machine, comme on disait alors. Mais les problèmes à résoudre à cette époque étaient encore mal connus, on s'orientait souvent vers un procédé de simple décryptage, lié comme cela se doit aux questions de fréquence, mais aussi au risque inhérent de considérer les langues comme de simples codages. Et pourtant, on le sait, les langues sont beaucoup plus complexes, ce sont des systèmes originaux, conditionnés par l'histoire, et qui réalisent des entités selon des universaux, mais aussi selon des valeurs propres. Les statistiques sont bien entendu fondamentales, et elles peuvent être une orientation vers des réponses potentielles. C'est d'ailleurs, à nouveau, à l'heure actuelle, une orientation de recherche très en vogue ; il est vrai que la constitution d'énormes bases de données textuelles fait « baigner » la recherche dans des exemples nombreux et réels et apporte des directions de réponses tout à fait indispensables dans l'exploitation des aspects statistiques probabilistes. On oublie trop souvent qu'en linguistique, comme dans tout autre domaine, de nombreux exemples appuyant une théorie ne prouvent pas grand-chose si la théorie peut être mise en cause par un seul exemple contraire. La linguistique, à cette époque,

n'avait pas toujours de solution satisfaisante. Un livre écrit en russe n'est pas simplement un livre écrit en français codé dans une autre langue ! Les unités linguistiques d'une langue à une autre ne sont pas toutes biunivoques. On ne traduit d'ailleurs pas des unités linguistiques, mais des unités cognitives ! L'accent n'était donc pas à mettre sur le décryptage, ni d'ailleurs uniquement et simplement sur la grammaire, comme on le faisait et comme on le fait malheureusement encore trop souvent. Il est vrai que la réexploitation des règles pour un autre projet de TA permet des économies substantielles. On ne peut donc négliger cet aspect, mais peut-être convient-il de rappeler que la syntaxe n'est qu'une série de règles déterminées par la sémantique et qu'à l'unité lexicale est attachée une prescription. Ce qui explique peut-être toutes les obligations de pré-édition et de post-édition pour certains systèmes. Les logiciels de traduction automatique qui ont déjà connu un succès évident sont ceux qui ont un sémantisme restreint et donc une syntaxe tout à fait contrôlable. La pierre d'achoppement de toute traduction est bien entendu le passage d'une série de cooccurrents où les unités linguistiques ont des « affinités », des « solidarités » dans une langue et qui correspondent ou non à d'autres collocations dans l'autre langue. Même la terminologie semble redécouvrir ce constat et modifie à l'heure actuelle son approche traditionnelle.

On trouvera dans ces Actes des textes originaux qui font le point des recherches, qui analysent les problèmes particuliers et qui proposent des méthodes d'évaluation. Aucune science ne peut se permettre d'ignorer cet aspect, la rétroaction est une des conditions de succès.

John Hutchins ouvre les Actes en présentant les nouvelles orientations des recherches depuis le début de notre décennie. Il met en lumière les orientations neuves et les innovations de la recherche et met l'accent tout particulier sur les développements récents, sur l'utilisation des systèmes de traduction automatique, sur les développements de la traduction automatique commerciale. Il termine son article en esquissant les perspectives que l'avenir nous réserve. Éva Dauphin met en relief les espoirs que suscitent les méthodes de recherche basées sur l'étude du corpus. Pour elle, la solution la plus efficace est la méthode d'étude du corpus qui répertorie et évalue les phénomènes particuliers dans un domaine spécifique. Cette solution permet d'extraire la terminologie particulière et de déterminer les phénomènes syntaxiques idoines, s'assurant ainsi d'une rentabilité plus certaine. Dominique Estival s'attache à présenter un projet de recherche dont l'objectif principal est l'élaboration d'une grammaire réutilisable pour l'analyse automatique. Il s'agit de « fabriquer » plusieurs boîtes à outils qui soient utilisables pour des applications variées. Liesbeth Degand s'attache à mettre en valeur la génération multilingue de textes. Elle examine les conditions lexico-grammaticales, sémantiques et ontologiques qui sont les composantes indubitables pour assurer un texte réel. Susan Warwick-Armstrong donne un vaste aperçu des problèmes qui se posent pour l'acquisition et l'exploitation de textes parallèles. Elle montre tout l'intérêt de tels ensembles pour la traduction. Gilles Sérasset et Étienne Blanc présentent leur recherche qui développe les bases lexicales multilingues par une approche interlingue s'appuyant sur les acceptions. Laurence Danlos analyse les multiples facettes de la traduction du verbe « faire » et dégage les classes de noms et les codages lexicaux qui en découlent. Christian Boitet met en lumière les choix à faire valoir en fonction des objectifs visés dans un système de TAO. Il analyse les diverses possibilités d'un système fondé sur le dialogue. Mutsuko Tomokiyo présente les recherches visant le transfert de la langue parlée dans un système de traduction auto-

matique. Elle analyse les règles de réécriture indispensables et leurs formalisations. Mathieu Lafourcade défend une solution générique qui règle la question des ambiguïtés dans le traitement automatique des langues naturelles. Alan Melby étudie le problème de la terminologie et sa réutilisation dans la traduction automatique. Il met en relief un légitime souci d'« efficience ». Igor A. Mel'čuk décrit le concept fondamental de fonction lexicale tel qu'il est illustré dans un dictionnaire explicatif et combinatoire et montre tout l'intérêt pour la traduction automatique. Alexandre Lazourski décrit un système de traduction automatique bidirectionnel qui s'appuie sur la théorie linguistique Sens-Texte de Mel'čuk. Sergei Nirenburg, David Farwell et Yorick Wilks s'attachent à examiner la théorie de traduction automatique basée sur la représentation des connaissances. Ils proposent l'introduction d'une architecture multimodulaire pour déterminer le meilleur choix du système de traduction. Deryle Lonsdale décrit les procédures qui permettent une optimisation de l'extraction des données lexicales pour servir à un système de traduction automatique. Siety Meijer analyse les problèmes créés au système de traduction automatique par les prépositions temporelles. Elle exploite à ce propos la théorie des fonctions lexicales de Mel'čuk. Adriane Rinsche présente une méthodologie pour l'évaluation des systèmes de traduction automatique et conclut par quelques recommandations pour un test comparatif des performances linguistiques des systèmes. Lidija Iordanskaja présente un modèle de génération de texte qui s'appuie sur la théorie Sens-Texte de Mel'čuk en utilisant pour la synthèse de la phrase une série de représentations sémantique et syntaxique profondes, puis de surface, et enfin phonétique et graphique. Gaston Gross étudie les connecteurs et l'orientation de la recherche pour le traitement automatique du langage. Marie-Claude L'Homme démontre un codage des phraséologismes verbaux pour résoudre les ambiguïtés traductionnelles. Graham Russell et Pierrette Bouillon présentent avec conviction les conditions qui doivent permettre de mieux déterminer un sous-langage et obtenir un traitement automatique plus facile mais aussi une plus grande fiabilité des résultats recherchés. Éric Wehrli revoit les problèmes que pose la traduction interactive et analyse, en fonction des destinataires potentiels, un système de traduction interactive retardée. Jacques Lerot explique la topicalisation et la focalisation dans un système de traduction automatique. Jean-Claude Lejosne avait la redoutable tâche de présenter la synthèse des questions étudiées et de déterminer la prospective. Il l'a fait avec une clarté indéniable.

Dans une séance parallèle, les intervenants ont décortiqué des questions liées plus globalement à la traduction. Ils ont notamment présenté pour discussion les divers aspects des études terminologiques. On lira l'étude de Christine Durieux qui oppose traductique et traduction humaine, celle de Pierre Lerat, qui évalue les modèles componentiels, casuels et relationnels dans le traitement automatique des langues naturelles. L'étude d'Yves Gentilhomme sur les termes et les symboles ouvre un vaste champ heuristique à l'étude sémiologique après avoir « décortiqué » avec une précision scientifique absolue les textes et discours technoscientifiques, obligeant ainsi à repenser concept « notion » et « concept ». Marc Van Campenhoudt définit la typologie des relations dans le cadre d'un projet de recherche portant sur l'élaboration des réseaux notionnels. Ingrid Meyer et Bruce McHaffie, en s'appuyant sur les derniers développements en génie cognitif, analysent les principes qui sous-tendent une approche basée sur la connaissance dans l'encodage des renseignements notionnels explicites. Élisabeth Blanchon, pour sa part, explore les « outils » logiciels disponibles qui rendent la traduction plus « conviviale ». Patrick Burkhard décrit la recherche terminologique à l'Union de Banques Suisses et Jean Quirion donne les orientations de la recherche terminotique au Gouvernement du Canada. Hussein Habaili propose les

principes directeurs qui ont été adoptés pour la création d'une banque de morphèmes-racines en arabe. Roger-Bruno Rabenilaina analyse les difficultés comparatives français-malgache à résoudre pour aboutir à un système de traitement automatique en traduction.

Le nombre de questions soulevées lors des présentations des communications montre encore que notre connaissance et notre compréhension des phénomènes linguistiques ont encore besoin de nouvelles explorations, d'idées fertiles, d'expériences encourageantes. Ces Actes, source d'information utile, cherchent à tracer une image de la recherche et à faire ainsi avancer la réflexion en créant des ententes salutaires entre les chercheurs qui ont un même but. La science est un partage, une interaction harmonieuse entre des chercheurs individuels qui mettent leur savoir en commun pour un meilleur environnement.

Pour terminer, on ne saurait oublier de dire merci à toutes celles et à tous ceux qui ont rendu ces Journées fructueuses et ces Actes possibles.

PARTIE I

1

Vers une nouvelle époque en traduction automatique

John HUTCHINS

Université d'East Anglia, Norwich, Angleterre

• *Abstract* •

Within the last few years, a number of new approaches and methods have changed the face of MT research. In the 1980s the dominant framework of MT was essentially 'rule-based' : e.g. the linguistics-based 'transfer' approaches of ARIANE, METAL, Eurotra, and Mu, the linguistics-based 'interlingua' models of Rosetta and DLT ; or the knowledge-based approach at Carnegie-Mellon.

Since 1990, a number of developments marked changes in the MT research picture, which may well indicate the beginning of a new era in the history of MT activity. Many are based on bilingual text corpora and employ statistical methods instead of analysis and generation rules and intermediary representations. In particular, there has been the IBM Candide project and experiments in example-based translation. Furthermore, there has been increasing interest in the problems of building large monolingual and bilingual lexical databases and of generating good quality output. And there have been other significant trends. In the past most systems were general-purpose ; now most are specialized in some respect. A feature of many recent systems is the restriction to controlled languages, to a sublanguage or to a specific domain, to a particular organization or to a particular user-type. At the same time, there are proposals for systems for users other than traditional translators.

In this paper I shall outline these recent developments in rule-based and corpus-based methods and touch briefly on interesting developments in commercial systems and in the application of systems. I shall end by attempting to forecast what the future may bring in MT research.

Introduction

Depuis environ 1989, la traduction automatique est entrée dans une période d'innovation méthodologique qui a changé l'optique de recherche.

Quel sont ces changements ? Quelle était la situation en traduction automatique il y a cinq ans ? Depuis 1975 jusqu'en 1988, on a vu apparaître un grand nombre de systèmes opérationnels et commerciaux notamment Systran, Logos, Météo, ainsi que plusieurs systèmes japonais. Ils utilisent, en général, la méthode *directe* de traduction ou la méthode *basée sur un transfert syntaxique,* et reposent sur des dictionnaires bilingues assez riches pour les domaines des textes à traduire ; l'analyse linguistique n'est pas très profonde ou reste abstraite, il n'y a presque pas d'analyse sémantique, ni d'exploitation de connaissances non linguistiques.

Quant à la recherche, on peut dire sans crainte de contradiction que, jusqu'à la fin des années 80, le cadre dominant a été l'approche basée sur les règles linguistiques : les règles d'analyse syntaxique, les règles lexicales, les règles pour la formation de représentations abstraites, les règles de désambiguïsation, les règles de transformation des arbres syntaxiques, les règles de transfert lexical, les règles de génération syntaxiques et morphologiques, etc. Vers 1985, on a commencé à développer des systèmes basés sur les connaissances du domaine des textes à traduire, mais cette approche est restée une nouveauté jusqu'à la fin de la décennie.

Depuis 1989, ce cadre dominant a été rompu par l'entrée en scène de méthodes et de stratégies nouvelles, appelées aujourd'hui les méthodes *basées sur corpus.* D'une part, un groupe à IBM a publié en 1989 les résultats de ses expériences avec un système de traduction purement statistique. L'efficacité de cette méthode a surpris beaucoup de chercheurs et a inspiré les expérimentations des années suivantes. D'autre part, des groupes japonais ont commencé à publier, à la même époque, leurs résultats préliminaires avec des méthodes *basées sur un corpus d'exemples* de traductions. Ces deux approches présentent une caractéristique commune : la comparaison de textes et le choix de traductions lexicales sans aucune règle syntaxique ou sémantique.

Dans cet article, nous nous concentrerons sur les nouveaux développements de la recherche et nous ne décrirons aucun projet en détail ; les systèmes cités ne sont que des exemples ; il en existe beaucoup d'autres (pour des références, voir mon étude récente – Hutchins 1993). Nous ne dirons presque rien des méthodes déjà bien établies à la fin des années 80. Nous ne parlerons pas non plus de l'utilisation des systèmes commerciaux, ni des aides automatisées pour traducteurs. Nous nous concentrerons donc exclusivement sur le développement des méthodes nouvelles en traduction automatique. Bien entendu, beaucoup de méthodes sont expérimentales et n'ont pas été mises à l'épreuve à grande échelle. Toutefois, les tendances que nous décrirons sont réelles ; la traduction automatique a connu récemment un renouveau dans sa méthodologie.

Les systèmes basés sur des règles

Avant de décrire ces nouveaux développements, nous commencerons par les approches basées sur les règles parce qu'il y a eu ici aussi des développements théoriques et méthodologiques d'une assez grande importance.

Il y a cinq ou six ans, se sont achevés deux grands projets basés sur l'approche de *transfert* : le projet Ariane à l'Université de Grenoble et le projet Eurotra des Communautés européennes. Ces systèmes illustraient les traits caractéristiques des sys-

tèmes dits de la *deuxième génération,* c'est-à-dire : les trois étapes d'analyse, de transfert et de synthèse ; les processus d'analyse et de génération qui utilisent plusieurs niveaux distincts de morphologie, de syntaxe et de sémantique ; des représentations d'interface assez abstraites en forme d'arbres étiquetés ; l'utilisation des règles pour transformer des arbres d'une étape à l'autre ; traitement par lots avec postédition et sans aucune intervention humaine au cours de la traduction ; et une absence presque totale d'informations pragmatiques et textuelles.

Cependant, les systèmes de transfert ont continué, comme en témoignent, par exemple, le système commercial Metal, le projet LMT d'une équipe IBM et le projet multilingue Eurolang, en cours de développement par la compagnie SITE en France avec la collaboration de la compagnie allemande Siemens-Nixdorf et qui se base sur les expériences du projet Eurotra.

L'approche *interlingue* a continué, elle aussi, avec encore plus d'intensité. Les traits distinctifs de cette approche sont bien connus : une langue pivot neutre pour représenter le sens des textes (l'interlangue), et des banques de connaissances dans le domaine des textes à traduire. À l'Université Carnegie Mellon, plusieurs modèles ont été développés, et 1992 vit l'inauguration d'un projet en collaboration avec la compagnie Caterpillar, qui vise à produire un système pour la traduction, sans postédition, des manuels techniques dans le domaine des engins de terrassement.

Il existe d'autres systèmes *interlingues*, par exemple le système ULTRA à l'Université de New Mexico et le système UNITRAN basé sur la théorie linguistique des principes et des paramètres. À cette liste s'ajoute le projet Pangloss, un système interlingue limité au sous-langage des fusionnements et des rachats, un projet en collaboration des universités de Southern California, New Mexico State et Carnegie Mellon, et qui utilise les expériences de ces équipes dans leurs propres recherches.

Les débuts de la tendance « lexicaliste »

Les systèmes basés sur les règles présentent un trait caractéristique : la transformation ou 'mappage' des représentations sous forme d'arbres étiquetés. Le système Eurotra, par exemple (figure 1), a proposé une série de transformations d'arbres : un arbre morphologique est transformé en un arbre syntaxique, un arbre syntaxique en un arbre sémantique, un arbre d'interface du texte source en un arbre équivalent du texte cible, etc. Essentiellement, un arbre doit satisfaire à des conditions précises, posséder une structure particulière et contenir des unités lexicales particulières ou des traits syntactiques ou sémantiques particuliers. De plus, les arbres eux-mêmes sont testés par les règles de formation – une grammaire vérifie la structure et les relations qui y sont représentées. Un arbre est rejeté s'il n'est pas conforme aux règles grammaticales du niveau en question : morphologique, syntaxique, sémantique, etc. Les grammaires et les règles de transformation déterminent les conditions ou contraintes qui limitent les possibilités de transfert d'un niveau à un autre et, en somme, d'un texte de la langue source à un texte de la langue cible.

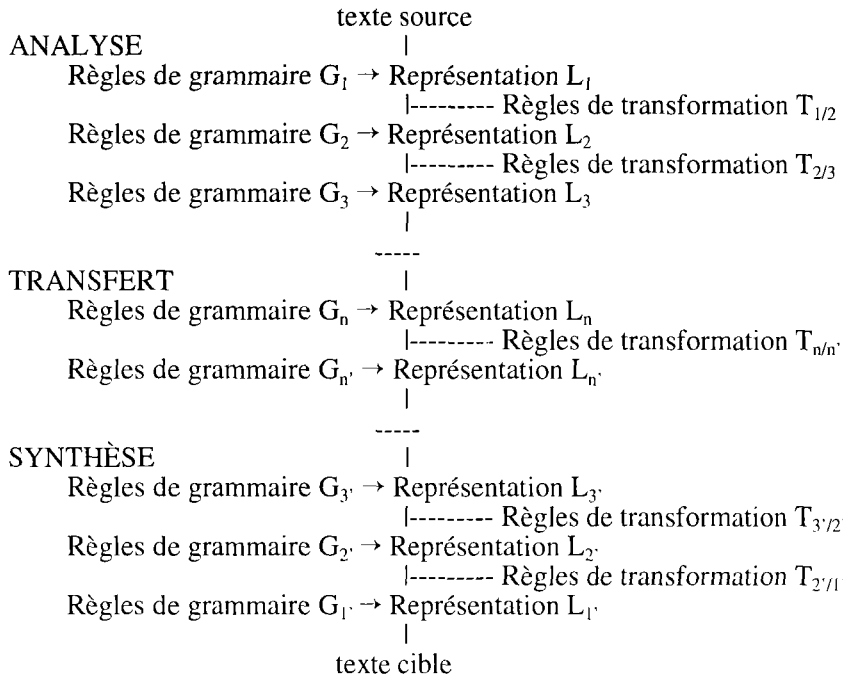


FIGURE 1 : Règles de formation et transformation (Eurotra).

Au cours des années se sont développés des formalismes basés sur les contraintes. Plutôt que d'utiliser un grand ensemble de règles qui ne peuvent s'appliquer que dans des circonstances et des représentations très particulières, on préfère un ensemble de règles abstraites assez restreint et on transfère les conditions et les contraintes dans les données lexicales. Par exemple (figure 2), pour traduire le verbe anglais *like* en verbe français *plaire,* il est nécessaire de transformer la structure syntaxique : le sujet anglais (*John*) devient un objet indirect en français (*à Jean*), et l'objet direct (*Mary*) devient un sujet en français (*Marie*). Ces conditions sont définies dans les ensembles de traits morphologiques, syntactiques et sémantiques des unités lexicales elles-mêmes. Un formalisme un peu plus complexe est nécessaire pour indiquer les contraintes attachées au mot anglais *likely* et au mot français *probable.* Le mot anglais exige un complément infinitif, tandis que le mot français exige une phrase subordonnée.

John likes Mary <--> Marie plaît à Jean

like, V :
(↑ PRED) = like <SUBJ, OBJ>
(τ↑ PRED FN) = plaire <SUBJ, OBJ>
(τ↑ AOBJ OBJ) = (SUBJ)
(τ↑ SUBJ) = (OBJ)

john, N :	*mary, N :*
(↑PRED) = john	(↑ PRED) = mary
(τ↑ PRED FN) = jean	(τ↑ PRED FN) = marie

F-structure de la langue cible :

$$\begin{bmatrix} \text{PRED} & \text{plaire} & \\ \text{SUBJ} & [\text{PRED marie }] & \\ \text{AOBJ} & [\text{OBJ} & [\text{PRED jean}] \,] \end{bmatrix}$$

Student is likely to work <--> Il est probable que l'étudiant travaillera

likely, A :	*probable, A :*
(↑ PRED) = likely <XCOMP> SUBJ	(↑ PRED) = probable <COMP>SUBJ
(↑ SUBJ) = (XCOMP SUBJ)	(↑ SUBJ FORM) = il
(τ↑ PRED FN) = probable	(↑ COMP COMPL) = que
(τ↑ COMP) = (XCOMP)	

F-structure de la langue cible :

$$\begin{bmatrix} \text{PRED} & \text{probable} & \\ \text{SUBJ} & [\text{FORM} & \text{il }] \\ \text{COMP} & \begin{bmatrix} \text{PRED} & \text{travailler} \\ \text{COMPL} & \text{que} \\ \text{SUBJ} & [...] \end{bmatrix} & \end{bmatrix}$$

FIGURE 2 : Formalisme basé sur contraintes (LFG).

Quant aux règles de transformation, elles sont devenues les méchanismes informatiques d'unification. Ce sont les règles d'unification qui dirigent l'interaction des ensembles de traits, la formation de nouveaux ensembles et l'élimination des ensembles illégitimes.

Face à l'orientation syntaxique qui a caractérisé beaucoup de systèmes de transfert dans le passé, on se tourne actuellement vers des solutions lexicales. Un exemple extrême de l'approche *lexicaliste* est la méthode qu'on appelle en anglais *shake and bake* (en français peut-être *agiter et cuire*) (figure 3).

unité lexicale monolingue (anglaise) :

[12]
$$\begin{bmatrix} \text{ORTHO } \textit{like} & \\ \text{SEM} \quad \text{E1 :} & \left\{ \begin{array}{l} \text{like (E1),} \\ \text{role (E1, experiencer, X1),} \\ \text{role (E1, stimulus, Y1)} \end{array} \right\} \\ \text{ARG0 E1} & \\ \text{ARG1 X1} & \\ \text{ARG2 Y1} & \end{bmatrix}$$

unité lexicale monolingue (espagnole) :

[13]
$$\begin{bmatrix} \text{ORTHO } \textit{gust-} \\ \text{SEM} \quad \text{E2} : \left\{ \begin{array}{l} \text{gustar (E2)} \\ \text{role (E2, stimulus, X2)} \\ \text{role (E2, experiencer, Y2)} \end{array} \right\} \\ \text{ARG0 E2} \\ \text{ARG1 X2} \\ \text{ARG2 Y2} \end{bmatrix}$$

entrée lexicale bilingue pour *like-gustar :*

$$\begin{bmatrix} \text{SPANISH} & [13] & \left[\text{SEM} \begin{bmatrix} \text{ARG0 E} \\ \text{ARG1 X} \\ \text{ARG2 Y} \end{bmatrix} \right] \\ \text{ENGLISH} & [12] & \left[\text{SEM} \begin{bmatrix} \text{ARG0 E} \\ \text{ARG1 Y} \\ \text{ARG2 X} \end{bmatrix} \right] \end{bmatrix}$$

FIGURE 3 : Approche lexicaliste (*Shake-and-bake*).

Il n'y a plus aucune représentation structurale ; il n'y a que des ensembles de représentations lexicales. Le processus de traduction exige l'identification des unités lexicales de la langue cible qui satisfassent les contraintes sémantiques, liées aux équivalents lexicaux de la langue source. Une traduction est formée (ou 'cuite') par les interactions entre les ensembles de traits et contraintes qui s'attachent aux mots de la langue cible.

Les origines des grammaires d'unification et des grammaires basées sur les contraintes remontent à presque dix ans. De nos jours, l'unification est devenue un concept central pour un grand nombre de théories linguistiques, et les grammaires et les formalismes basés sur les contraintes attirent beaucoup de chercheurs en traduction automatique : LFG (Grammaire fonctionnelle lexicale), Grammaire à clauses définies, GPSG, Grammaire catégorielle, etc. (ainsi que la grammaire d'UNITRAN basée sur les principes et les paramètres). L'avantage majeur de ces grammaires est la simplification des règles d'analyse, de transformation et de génération. À une série de représentations complexes d'information à plusieurs niveaux, on préfère des représentations monostratales ou même le transfert au moyen d'unités lexicales simples. En même temps, les composantes de ces grammaires sont, en principe, réversibles. Il n'est pas nécessaire de construire pour la même langue différentes grammaires d'analyse et de génération ; le même formalisme et les mêmes grammaires peuvent être appliqués dans les deux sens.

Plusieurs équipes ont construit des systèmes généraux pour le traitement de la langue naturelle, basés sur les grammaires d'unification et des contraintes, qui s'appliquent aussi à la traduction. Le système CLE (*Cove Language Engine*), par exemple, a été appliqué à la traduction automatique du suédois à l'anglais et de l'anglais au suédois ; tandis que le système PLNLP a fourni la base de systèmes de traduction pour le portugais, le coréen et le japonais. Mieux connu dans le domaine de la traduction automatique est l'Environnement Linguistique d'Unification (ELU) développé par une

équipe suisse à Genève. Sur cette base, a été développé un système bidirectionnel pour la traduction des bulletins d'avalanches entre l'allemand et le français.

Les lexiques et la génération de textes

La tendance vers les approches lexicalistes a eu deux conséquences : l'une pour les lexiques et l'autre pour la génération de textes. Pour les lexiques on observe une augmentation dans l'étendue des informations liées aux unités lexicales. Il ne s'agit plus seulement de données morphologiques et grammaticales des mots de la langue source et des mots et phrases correspondants du langage cible. On a ajouté les contraintes syntaxiques et sémantiques et des informations non linguistiques et conceptuelles, souvent limitées au domaine particulier des textes à traduire, comme en témoignent en particulier les systèmes basés sur une interlangue, par exemple dans les systèmes développés à Carnegie Mellon ou dans le système UNITRAN.

On voit croître l'intérêt porté aux problèmes de la construction des lexiques pour la traduction automatique. La construction d'un lexique, qui soit suffisamment grand pour les applications réelles et pratiques d'un système opérationnel, est difficile et coûteuse. Beaucoup d'équipes cherchent des méthodes d'acquisition des informations lexicales, qui utilisent des ressources lexicographiques déjà disponibles, par exemple des dictionnaires bilingues destinés aux étudiants ou consacrés aux différents domaines scientifiques. De même, les groupes de recherche collaborent de plus en plus étroitement pour la construction de lexiques embrassant une grande gamme d'applications et une variété importante de systèmes. L'exemple le plus connu est sans doute le projet collaboratif EDR (*Electronic Dictionary Research*) de plusieurs compagnies japonaises.

Quant à la génération, il y a dix ans encore, on croyait que les composantes les plus problématiques étaient celles de l'analyse syntaxique et sémantique, de la désambiguïsation du sens et de l'identification des antécédents des pronoms – en somme, de celles de la compréhension du texte à traduire. À cette époque, on négligeait généralement les problèmes de la génération de textes idiomatiques dans la langue cible. Aujourd'hui, c'est évident que, pour améliorer la qualité des traductions automatiques, il faut non seulement faire attention à l'analyse, mais aussi à la sélection des mots propres et à l'idiomatisme des textes produits. La génération de textes n'est donc plus négligée en traduction automatique.

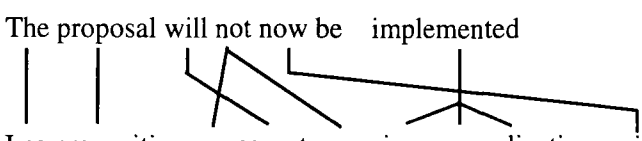
Les systèmes basés sur un corpus de textes bilingues

Ces deux tendances ont été accentuées fortement par les développements de nouvelles méthodes qui se basent sur des corpus de textes bilingues, en particulier les méthodes statistiques et les méthodes basées sur des exemples de traductions.

Le développement le plus étonnant a été le projet Candide d'un groupe de chercheurs à IBM. Sa méthode se fonde sur l'analyse stochastique de l'énorme corpus de textes, anglais et français, constitué de rapports des débats parlementaires du Canada. La première tâche est d'aligner les phrases, les groupes de mots et les mots individuels des textes parallèles, une étape accomplie sans aucune information linguistique. On voit, par exemple (figure 4), que cette méthode met en parallèle *proposal* et *proposition*, *now* et *main-*

tenant, et le mot anglais *implemented* et la phrase française *mises en application.* Mais, contrairement à toute intuition linguistique, elle aligne *will* et *seront*, tandis que pour le mot *be* aucun mot français n'est mis en parallèle. À partir d'un grand nombre d'alignements anglais-français, on calcule les fréquences et les probabilités des correspondances. Le mot anglais *not* correspond le plus souvent à deux mots français (fertilité 2 avec une probabilité de 0.758) et ces deux mots sont en général *ne* et *pas* (probabilités de 0.469 et 0.460) ; les autres correspondances sont moins probables : *non* (0.024), *pas du tout* (0.003), etc. La méthode a été évaluée pour la traduction de l'anglais au français.

Alignement :



Anglais : *not*

Français	Probabilité	Fertilité	Probabilité
pas	.469	2	.758
ne	.460	0	.133
non	.024	1	.106
pas du tout	.003		
faux	.003		
plus	.002		
etc.			

FIGURE 4 : TA stochastique (Candide, IBM).

Pour beaucoup de chercheurs, le plus surprenant est la qualité des résultats : presque la moitié des phrases traduites correspondaient exactement aux traductions dans le corpus, ou bien le même sens était exprimé avec des mots un peu différents, ou encore d'autres traductions légitimes étaient offertes. Pour améliorer ces résultats, les chercheurs d'IBM ont proposé d'introduire des méthodes statistiques plus complexes, ainsi que quelques éléments minimaux d'information linguistique – ils sont empiristes et prêts à expérimenter avec d'autres moyens. Ils proposent d'utiliser l'information morphologique, par exemple, de traiter toutes les variantes d'un verbe comme un seul mot, et d'utiliser certaines transformations syntaxiques, par exemple, la transformation de structures discontinues en structures qui ressemblent plus exactement à celles de la langue cible (figure 5).

Has the store any eggs? -> The store has any eggs QINV
John does not like turnips -> John likes do_not_M1 turnips
Je ne sais pas -> Je sais ne_pas
Je vous le donnerai -> Je donnerai le_DPRO vous_IPRO

FIGURE 5 : Transformations syntaxiques (proposées).

La méthode basée sur des exemples de traduction a profité, elle aussi, du développement de logiciels pour l'accès rapide aux banques de données textuelles. Pour cette méthode, il existe une banque de textes bilingues en parallèle, qui ont été alignés, soit par les méthodes statistiques, soit par l'analyse morphologique et syntaxique. L'essence de la méthode est l'extraction et la sélection des phrases équivalentes. Si, par exemple (figure 6), on cherche une traduction pour le mot anglais *fields*, on trouverait peut-être dans une banque de données les possibilités françaises suivantes : *domaines, activités, champs*. Pour chaque exemple on trouve aussi des contextes. S'il y a une correspondance exacte *(coal fields -> bassins houillers),* la sélection est immédiate. Mais, s'il n'y a pas une correspondance exacte, il faut utiliser des algorithmes pour trouver l'équivalent propre.

anglais	français
the main fields	les principaux domaines
the following fields	les domaines suivants
these two fields	ces deux domaines
the specialized fields	les domaines spécialisés
the para-medical fields	activités paramédicales
the magnetic fields	les champs magnétiques
the coal fields	les bassins houillers
the corn fields	les champs de blé

FIGURE 6 : Banque d'exemples de traductions : *field.*

Quelques équipes utilisent les méthodes sémantiques, par exemple un réseau sémantique ou une hiérarchie de termes d'un domaine. D'autres équipes utilisent les informations statistiques à propos des fréquences lexicales dans la langue cible. Généralement, la méthode des exemples est utilisée comme complément aux méthodes plus traditionnelles basées sur les règles linguistiques. Puisque les textes sont des exemples de traductions humaines, il est certain que les résultats seront aussi idiomatiques que possible. Il est difficile, quand on traduit du français vers l'anglais, de choisir la préposition exacte qui correspond au petit mot *de* (figure 7). Une banque de données qui contient un grand nombre d'exemples peut être très utile pour la sélection des équivalents. Elle l'est encore plus face à des difficultés plus grandes, par exemple, celles que représente la traduction française de la phrase anglaise *have an effect on* (figure 8).

français	anglais
le livre *de* mon père	*my father's book*
un verre *d'*eau	*glass of water*
il est certain *de* réussir	*certain to succeed*
il est capable *de* résister	*capable of resisting*
il vient *de* Paris	*he comes from Paris*
le train *de* Paris	*the train to/from Paris*
il partit *de* nuit	*he left at night*
il partit *de* bonne heure	*he left in good time*
je suis âgé *de* trente ans	*I am thirty years old*

FIGURE 7 : *de.*

anglais	français
have a direct effect on	ont une influence directe à
have a direct effect on	intéressent directement
have a direct effect on	ont eu une répercussion directe sur
has had a marked effect on	a largement influencé
had a positive effect on	s'est avérée positive dans
had a highly negative effect on [X]	[X] en auraient été gravement affectés
will have a decisive effect on	influencera de façon déterminante
would have a detrimental effect on	aurait de fâcheuses répercussions sur

FIGURE 8 : *...have a/an effect on...*

Une banque de textes bilingues en parallèle peut être également utilisée dans d'autres buts. Plusieurs équipes mènent des expériences avec des stations de travail destinées aux traducteurs humains. L'un des groupes les plus actifs dans les méthodes statistiques pour aligner les textes bilingues (c'est-à-dire le groupe à AT&T Bell) prévoit l'application d'un corpus aligné comme base de connaissances pour les traducteurs. Un autre groupe très actif est l'équipe canadienne bien connue du Centre d'innovation en technologies de l'information. Son but est le développement d'outils pour traducteurs, qui permettent, entre autres, de chercher une banque de textes pour les exemples d'utilisation de n'importe quel mot anglais ou français dans un contexte particulier.

La méthode des exemples de textes a renforcé aussi la tendance que nous avons déjà mentionnée à propos de la recherche de génération de textes. En dehors de la traduction automatique, des groupes s'intéressent de plus en plus à la représentation des données informatiques dans une langue naturelle idiomatique. La génération de textes dans plusieurs langues a été le but de deux projets canadiens : dans le projet RAREAS, il s'agit de produire les textes anglais et français des prévisions maritimes ; dans le projet LFS, des sommaires bilingues sur le marché du travail.

Un autre stimulant, qui illustre lui-même une tendance importante de ces cinq dernières années, c'est qu'on a reconnu une demande pour des types de traduction jamais étudiés auparavant. Dans le passé, les systèmes ont été, en général, construits pour des personnes bilingues, pour les traducteurs et pour ceux qui connaissent les langues source et cible. En outre, les textes traduits exigeaient une postédition. On a négligé les besoins de ceux qui ne connaissent pas la langue cible. Ce sont souvent des négociants et des hommes d'affaires qui font du commerce à l'étranger et qui veulent communiquer un message assez simple dans une langue inconnue. Récemment, quelques équipes ont fait des expériences avec des systèmes où le texte à traduire est le fruit d'une collaboration entre homme et ordinateur (par exemple, UMIST et l'Université de Bruxelles). Il est ainsi possible d'établir un texte que le système est capable de traduire sans se référer davantage à l'auteur, qui n'exige aucune révision et dont la qualité de traduction est bien assurée.

Les systèmes destinés à des utilisateurs particuliers

Il y a toujours eu des systèmes de traduction automatique limités à des domaines res-

treints. Même les systèmes destinés à l'usage général se limitent, en fait, à quelques domaines particuliers. De cette façon, il est, en effet, possible de construire un lexique suffisant et de réduire les problèmes d'ambiguïté et de choix des équivalents dans la langue cible en restreignant le vocabulaire et les structures grammaticales des textes à traduire. Bien que les frais de l'édition préalable puissent être assez élevés, la postédition est ainsi réduite de façon considérable. On peut aussi limiter le système lui-même à un sous-langage spécifique : Météo, par exemple, qui traduit, depuis 15 ans, les rapports météorologiques. Parmi les systèmes pour sous-langages et langues restreintes, on compte, ces dernière années, le système CRITTER, pour les rapports du marché des bestiaux, des projets déjà mentionnés (ELU, Pangloss et Caterpillar) et les projets très ambitieux destinés au développement de systèmes de traduction de la langue parlée. Le projet ATR au Japon dure déjà depuis sept ans et continuera jusqu'à la fin du siècle ; c'est un système pour les renseignements sur les conférences internationales et pour les enregistrements téléphoniques dans les hôtels. Le projet VERBMOBIL en Europe vise à développer une aide portative dans les négociations commerciales face à face conduites en anglais par les Allemands ou les Japonais qui ne connaissent pas très bien la langue anglaise.

Dans le passé, peu de systèmes ont été construits par les utilisateurs eux-mêmes. La PAHO (Organisation Pan-Amérique de la Santé), a cependant développé deux systèmes pour la traduction de l'anglais à l'espagnol et de l'espagnol à l'anglais. Au cours de ces dernières années, plusieurs systèmes de ce type ont été mis en œuvre. Ce sont typiquement des systèmes avec des vocabulaires restreints pour un domaine particulier et basés sur un sous-langage spécifique. C'est un signe encourageant que les méthodes informatiques de la traduction automatique et du traitement de la langue naturelle se soient maintenant répandues de plus en plus au delà des cercles limités de chercheurs. De tels systèmes ne sont pas innovatifs au point de vue théorique et méthodologique, mais ils sont souvent très avancés en ce qui concerne la technologie informatique. Je crois que c'est une tendance que nous verrons se répandre rapidement au cours des années à venir.

Une nouvelle époque ?

À notre avis, la recherche en traduction automatique a traversé cinq époques jusqu'à nos jours. La première a commencé avec le mémorandum de William Weaver, en 1949, qui a lancé les recherches. La deuxième a commencé avec la démonstration, en 1954, d'un système assez simple pour la traduction du russe à l'anglais, qui a encouragé les agences gouvernementales des États-Unis et d'autres pays à soutenir les recherches à une grande échelle. Elle s'est terminée avec le fameux rapport d'ALPAC. La troisième époque a duré jusqu'à environ 1975, date d'un nouvel essor grâce à l'intérêt croissant au Canada et en Europe. Alors que les systèmes des deux premières époques ont été, en général, basés sur la méthode directe de traduction, depuis ALPAC, la plupart des systèmes utilisent la méthode du transfert ou interlingue et sont basés sur les règles. Mais, comme nous l'avons déjà mentionné, il existe, à présent, de nouvelles méthodes et de nouvelles tendances : les approches basées sur les corpus bilingues de textes, les méthodes statistiques et les méthodes basées sur des exemples, ainsi que de nouvelles méthodes basées sur les grammaires d'unification et de contraintes. Ces innovations sont apparues pendant les cinq dernières années et la recherche en traduction automatique me semble être entrée dans une nouvelle époque.

Si la méthode directe a caractérisé la *première génération* et les méthodes indirectes de transfert et interlingue la *deuxième génération*, il faut se demander quelles seront les caractéristiques potentielles de la future *troisième génération*.

Les systèmes de « troisième génération » ?

De l'avis de nombreux spécialistes, les systèmes futurs combineront des méthodes basées sur les règles assez traditionnelles, avec des méthodes statistiques et basées sur les exemples. Ils seront hybrides. De quelle façon ? On peut imaginer que les méthodes linguistiques des systèmes indirects fourniront la base pour l'application des méthodes basées sur les banques de connaissances, sur les données statistiques et sur les exemples de textes traduits. La base linguistique fournira des analyses syntaxiques et sémantiques assez simples, l'essentiel du transfert et les informations syntaxiques pour la génération. Il s'agira probablement de formalismes d'unification et de grammaires basées sur les contraintes. Les autres méthodes permettront une désambiguïsation plus souple et fourniront les informations nécessaires pour le transfert lexical et pour la production de textes idiomatiques.

Ainsi, en ce qui concerne la base de règles linguistiques, on peut prévoir que :

- les règles seront moins ambitieuses que celles des systèmes indirects de transfert et d'interlingue ;
- l'analyse syntaxique sera limitée à la reconnaissance des structures superficielles, des composantes de phrase et des relations de dépendance ;
- il n'y aura presque aucune analyse profonde des relations logiques ;
- l'analyse sémantique sera limitée à l'identification des rôles des éléments phrasaux : l'agent, l'instrument, etc. ;
- les informations lexicales seront extraites principalement de ressources régulières comme des dictionnaires généraux destinés au grand public ; par conséquent, elles n'indiqueront que des catégories syntaxiques et peut-être des traits sémantiques peu raffinés ;
- ces traits sémantiques assez rudimentaires ne seront utilisés que pour une désambiguïsation initiale ;
- les règles de transfert lexical et structural s'appliqueront peut-être à des représentations peu profondes (quoique moins simples que dans le système d'IBM).

Quant aux méthodes encore assez nouvelles :

- des exemples de traduction, stockés dans une banque de textes bilingues alignés, seront utilisés pour aider la désambiguïsation plus délicate dans l'analyse de la langue source et pour choisir les équivalents dans la langue cible ;
- des données statistiques sur les collocations lexicales et les fréquences du vocabulaire monolingues aideront l'analyse syntaxique et sémantique des phrases, la désambiguïsation monolingue, et le choix des phrases idiomatiques dans la langue cible ;
- des données sur les probabilités des équivalences bilingues seront utilisées pendant le transfert lexical ;
- des banques de connaissances des domaines en question aideront la désambiguïsation monolingue et interlingue.

On peut anticiper aussi d'autres développements importants :

- l'emploi des méthodes de *feedback* pour améliorer les grammaires (ou le fond de règles) et les lexiques monolingues et bilingues ;
- les recherches plus approfondies sur les problèmes de discours et de style ;
- l'intégration de la traduction automatique dans les systèmes pour la production, la transmission et la gestion de documents dans les bureaux (à l'exemple des postes de travail pour traducteurs).

L'utilisation des systèmes

Aujourd'hui, les utilisateurs et les chercheurs sont tous réalistes. Un système de traduction complètement automatique qui produirait des textes idiomatiques comparables aux traductions humaines n'est que le rêve de ceux qui n'en ont pas l'expérience. On ne croit pas non plus à la réalisation pratique et économique de systèmes généraux, qui traitent une grande gamme de domaines. La recherche est centrée sur le développement de systèmes limités à un sous-langage ou à un domaine technique spécial. Dans des conditions propices, les systèmes, loin d'être parfaits, peuvent être utilisés avec profit et succès. Tout le monde, sans doute, aspire à des systèmes de meilleure qualité, mais on ne les attend pas dans un proche avenir. Il faut rappeler que ce sont toujours les systèmes expérimentaux qui expérimenteront des techniques nouvelles. On n'attend pas, avant la fin du siècle, l'arrivée d'un système commercial basé sur les méthodes de la *troisième génération.* À cette nouvelle époque, les systèmes commerciaux seront basés sur des méthodes sûres, bien établies et bien éprouvées – qui toutefois ne seront pas les méthodes les plus innovatrices.

Néanmoins, on doit prévoir une expansion rapide du nombre d'utilisateurs des systèmes de traduction automatique. Dans les dernières années, le nombre de pages traduites automatiquement a crû fortement – à présent, plus d'un million de pages par année, ou trois cents millions de mots (Vasconcellos 1993). L'expansion a eu lieu dans les grandes compagnies multinationales et dans les agences de traduction, surtout pour la traduction des manuels techniques. Mais il y a eu aussi une expansion du nombre des utilisateurs non professionnels. Beaucoup d'entre eux ont acquis des systèmes bon marché pour les ordinateurs personnels – qui sont, bien sûr, des systèmes assez simples quant aux méthodes linguistiques et qui n'ont pas été développés par les experts en traduction automatique. On peut douter de l'efficacité des systèmes, mais les besoins de ces utilisateurs sont incontestables. C'est un marché que peu de chercheurs spécialistes ont considéré à ce jour.

En somme, on prévoit une période d'expérimentation variée et assez incohérente quant à la théorie. On prévoit aussi une expansion de l'utilisation des systèmes de grande taille ainsi que du nombre d'utilisateurs de systèmes basés sur les ordinateurs personnels. En particulier, on prévoit une expansion dans le nombre d'utilisateurs par le moyen de réseaux électroniques ; en France et au Japon, la traduction automatique a été déjà offerte sur les réseaux PC-VAN, Niftyserve et Minitel ; et, cette année, CompuServe annoncera bientôt un service de traduction automatique. Quelles sortes de systèmes satisferont ces nouveaux besoins ? Les chercheurs devront relever ces défis, et aussi, s'attacher avec d'autant plus d'urgence, à établir des normes qui permettent d'évaluer et de comparer la performance, la qualité et l'efficacité des systèmes commerciaux. Dans le proche avenir, on attend, avec beaucoup d'intérêt et d'espoir, des changements rapides dans le domaine de la traduction automatique.

Références

Les sources principales de cet article sont les comptes rendus des congrès portant sur la traduction automatique : TMI-90, TMI-92, TMI-93, MT Summit III et MT Summit IV. Voir, notamment, Hutchins (1993) pour des références plus précises. Pour des articles en français, voir aussi le numéro spécial de la revue *Meta*, vol. 37 n° 4, décembre 1992.

TMI-90. *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* TMI-90, Linguistic Research Center, University of Texas at Austin, États-Unis, 11-13 June 1990.

TMI-92. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* TMI-92, Empiricist vs Rationalist Methods in MT, CCRIT, Montréal, Canada, June 25-27 1992.

TMI-93. *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. MT in the Next Generation,* Kyoto, Japon, July 14-16 1993.

Proceedings of MT Summit III, Washington DC, États-Unis, July 1-4 1991.

Proceedings of MT Summit IV : International Cooperation for Global Communication, Kobe, Japon, July 20-22 1993.

HUTCHINS, W. J. (1993) : « Latest Developments in Machine Translation Technology », *Proceedings of MT Summit IV,* pp. 11-34.

VASCONCELLOS, M. (1993) : « The Present State of Machine Translation Usage Technology, or : How Do I Use Thee? Let Me Count the Ways! », *Proceedings of MT Summit IV,* pp. 35-46.

2

Étude de corpus : un préalable pour l'adaptation des systèmes de traduction automatique aux besoins des utilisateurs

Éva DAUPHIN

AÉROSPATIALE, Centre commun de recherches Louis-Blériot, Suresnes, France

• *Abstract* •

This paper intends to show the interest of the corpora studies for a potential user of MT either evaluating and enhancing an existing system or specifying a new one. The theoretical method that AÉROSPATIALE has chosen in cooperation with EDF and GSI-ERLI is then presented step by step as well as the changes implied by its partial automation performed by these three French partners of the EUREKA project GRAAL (Reusable Grammars for the Automatic Analysis of Languages).

Introduction

Les études de corpus sont apparues nécessaires à l'AÉROSPATIALE après plusieurs années d'expérimentation de systèmes de traduction automatique (TA) existant sur le marché et d'évaluation de leurs performances. L'étude de corpus constitue pour nous un moyen d'identifier les besoins et contraintes linguistiques qui doivent être prises en compte par un consommateur potentiel de traduction automatique. Le besoin de l'utilisateur correspond, dans cet article, à la qualité linguistique minimale de la traduction obtenue, les contraintes à la qualité linguistique des textes qu'il désire traiter.

Remarque : Dans ce contexte, on entend par corpus un ensemble de textes homogènes, c'est-à-dire traitant du même domaine, rédigés et utilisés par le même type

de personnes et dans des conditions semblables (par exemple : l'ensemble des manuels de maintenance hélicoptère émanant de la société AÉROSPATIALE).

Pourquoi réaliser une étude de corpus ?

L'identification des besoins et contraintes linguistiques est primordiale pour l'utilisateur dans deux situations possibles :

- évaluation et adaptation d'un système existant ;
- participation active à l'élaboration d'un nouveau système.

Évaluation et adaptation d'un système existant

Comme pour tout produit, l'évaluation des performances linguistiques d'un système de TA doit être réalisée en fonction des besoins et contraintes de l'utilisateur. L'évaluation consiste, ici, à vérifier l'aptitude d'un système à traiter les textes propres à l'utilisateur et à en donner une traduction correspondant aux critères établis par ce dernier.

À l'heure actuelle, l'utilisateur dispose de deux méthodes pour évaluer les performances linguistiques d'un système :

- traduire des textes et analyser les erreurs de traduction commises par le système ;
- élaborer des séries de tests aléatoires pour connaître la couverture linguistique du système.

L'analyse d'erreurs est un travail long et fastidieux qui oblige à passer en revue l'ensemble de la traduction en se référant au texte source et éventuellement à sa traduction humaine validée. Cette opération ne peut être réalisée que manuellement et sa faisabilité dépend donc du volume des textes étudiés. Cette méthode est pourtant la seule existante qui permet de vérifier les performances d'un système en situation réelle. En outre, l'analyse d'erreurs pose des problèmes méthodologiques sérieux en termes de critères de classement (par partie du discours, par type de correction, selon l'origine de l'erreur, etc.).

L'étude de corpus intervient ici comme un moyen de recenser et de modéliser les caractéristiques linguistiques des textes permettant d'élaborer des séries de tests portant sur les phénomènes représentatifs d'un corpus, réduisant ainsi les volumes de données à analyser. L'analyse s'en trouve facilitée grâce à l'isolement de chacun des phénomènes. En effet, les expériences d'analyse d'erreurs menées à AÉROSPATIALE montrent que la plupart des segments *mal traduits* cumulent plusieurs erreurs et il est alors difficile de déterminer les phénomènes mis en cause. Par exemple :

Light, medium and heavy-lift helicopters

est traduit :

La lumière, moyenne et lourd-soulèvent des hélicoptères

L'étude de corpus représente ici un moyen de modéliser ces cumuls de phé-

nomènes en les répertoriant au préalable pour aborder ce problème de façon plus progressive :

1. problème de l'homographie *adj/nom* ;
2. problème des adjectifs composés en *adj-participe* ;
3. problème des adjectifs en énumération ;
4. problème de restitution de la place des adjectifs.

Ces quatre problèmes peuvent être testés de façon isolée dans un premier temps, puis combinés de façon progressive afin de déceler des interférences de règles possibles.

L'autre intérêt d'élaborer des phrases tests à partir des résultats de l'étude de corpus est de limiter l'étendue des tests aux phénomènes linguistiques couverts par les textes. Les phénomènes absents du texte ne seront pas testés, mais les formes identifiées, non prévues par les grammaires traditionnelles, le seront, réduisant ainsi le champ de l'évaluation aux seuls corpus envisagés. En effet, la deuxième méthode d'évaluation des performances d'un système, plus empirique, consiste à élaborer des fichiers de tests pour estimer la couverture linguistique du système. Elle implique de passer en revue, sous forme de phrases tests, l'ensemble des phénomènes de la langue, posant le problème important de l'exhaustivité. En effet, en admettant que l'ensemble des phénomènes de la langue soit représenté dans les tests, qu'en est-il des phénomènes agrammaticaux récurrents existant dans les corpus ?

L'intervention des résultats d'une étude corpus dans le processus d'évaluation permet ainsi de bénéficier des avantages des deux méthodes d'évaluation citées et d'en limiter les inconvénients :

- La **notion de représentativité** des tests par rapport au corpus est préservée permettant, entre autres, d'estimer le taux de couverture lexicale du système par rapport à la réalité du vocabulaire employé dans le corpus. Cette opération est essentielle si l'utilisateur désire évaluer l'importance de l'alimentation lexico-terminologique qui sera nécessaire pour traiter le corpus.

- L'**aspect de modélisation** des phénomènes pour l'élaboration de phrases tests autorise une diminution des volumes à analyser et une démarche progressive d'identification du phénomène litigieux.

- La **limitation des tests** aux seuls phénomènes présents dans le corpus restreint leur nombre et surtout permet une évaluation ciblée des possibilités du système pour un corpus particulier en déterminant la couverture linguistique minimale que le système doit présenter pour ce corpus.

En terme d'amélioration du système, donc de son adaptation aux réels besoins de l'utilisateur, l'étude de corpus apparaît comme un moyen d'évaluer plus efficacement les possibilités d'un système pour un corpus donné et un moyen de mieux isoler et formaliser les améliorations qui devront être réalisées.

L'étude de corpus intervient, ici, comme le moyen pour l'utilisateur d'identifier les contraintes linguistiques (qualité du texte source) qu'il doit prendre en compte pour une évaluation personnalisée.

Participation active à l'élaboration d'un nouveau système

Cette situation est, bien entendu, idéale pour l'utilisateur qui désire disposer d'un outil complètement adapté à ses besoins. Elle implique pour cet utilisateur de définir précisément ses contraintes linguistiques et ses objectifs. Pour le système, cette démarche signifie qu'il devra présenter une couverture linguistique minimale correspondant aux caractéristiques des corpus envisagés. Seule une étude de corpus minutieuse permet d'identifier l'ensemble des phénomènes linguistiques qui doivent impérativement être traités conformément aux besoins de l'utilisateur pour qu'il accepte réellement le système.

Dans le cadre de la participation aux spécifications des besoins et la définition des contraintes linguistiques, il est nécessaire de procéder en plusieurs étapes :

- identification des phénomènes linguistiques présents dans le corpus envisagé ;
- formalisation de ces phénomènes ;
- intégration des spécifications dans les composantes du système.

L'étude de corpus constitue le moyen de réaliser la première étape en recensant et en modélisant les phénomènes linguistiques caractéristiques d'un corpus. Cet exercice permet à l'utilisateur d'élaborer des spécifications linguistiques qui seront ensuite soumises aux développeurs de l'application, leur permettant ainsi d'organiser le contenu et le fonctionnement des différents modules du système pour adapter celui-ci au corpus traité. En effet, les résultats d'une étude de corpus se situent à des niveaux différents et leur intégration dans un système concerne tant les grammaires, les ressources lexicales, terminologiques que les modules de prétraitement, de post-traitement, etc.

Étude de corpus : la démarche méthodologique

La vocation d'une étude de corpus étant de recenser le plus d'informations linguistiques possibles, la démarche préconisée se veut des plus exhaustives bien que la réalisation de certains des points d'étude proposés puisse sembler complexe.

La méthode initiale est constituée de trois parties comportant chacune un ensemble de points qui doivent être étudiés de façon chronologique, certaines observations pouvant en effet constituer le point de départ d'autres observations. Toutefois, nous verrons que l'application effective et l'automatisation partielle de cette méthode a engendré certains remaniements et en particulier, l'ajout d'une partie *formats*.

L'objectif étant de passer en revue tous les aspects linguistiques des textes, nous avons organisé l'étude en trois parties auxquelles nous avons ajouté un volet *formats* qui s'est avéré nécessaire pour l'automatisation des recherches.

Étude lexicale et terminologique des textes

Cette partie consiste à isoler l'ensemble des lexies du corpus par partie du discours afin d'évaluer la couverture lexicale du corpus dans un premier temps. Une attention

particulière est accordée ensuite à l'identification des sigles, des noms propres et des mots en langue étrangère, souvent problématiques en TA, parmi les substantifs.

Dans un deuxième temps, un rapprochement des listes est réalisé avec les ressources lexicales existantes afin d'évaluer la part du vocabulaire du corpus inconnu dans les ressources lexicales. La consultation de la liste des mots *inconnus* ainsi obtenue permet déjà de calculer la part d'erreurs orthographiques (et leur importance en nombre) ayant entraîné la non-reconnaissance du mot par le dictionnaire.

Dans un troisième temps, un recensement des termes potentiels est effectué à l'aide de *patrons* (*patterns*) prédéfinis du type N + N, N à N, N de N, etc. Le rapprochement de la liste ainsi obtenue avec les ressources terminologiques existantes permet d'évaluer le codage terminologique qui doit être réalisé et permet en outre, de mieux cerner le ou les domaine(s) terminologique(s) dominant(s) du corpus.

Remarque : Dans cet article, une distinction claire est faite entre les ressources dites lexicales composées du vocabulaire général de la langue et les ressources terminologiques qui contiennent, elles, le vocabulaire spécifique d'AÉROSPATIALE tant au niveau de ses activités de production qu'au niveau de ses produits.

L'étude lexicale et terminologique détaillée du texte constitue une base de réflexion en ce qui concerne l'organisation des ressources nécessaires et introduit le problème de l'homographie qui constitue une des difficultés rencontrées en TA. Assortie de calculs statistiques, cette partie de l'étude de corpus peut permettre d'évaluer le degré de répétition du texte, critère non négligeable quand on aborde la traduction de corpus très volumineux.

Étude morpho-syntaxique des textes

C'est au cours de cette étude que les règles syntaxiques auxquelles obéit le corpus doivent être identifiées. Pour ce faire, une analyse des phrases doit être réalisée dans un premier temps, afin d'en identifier la structure. Une fois la structure des phrases identifiée, il semble plus aisé de recenser le type de propositions utilisées et d'estimer le degré de complexité des phrases.

L'étude des propositions est essentielle si l'on veut recenser l'ensemble des conjonctions et relateurs dont le comportement doit être connu pour en assurer un traitement correct par le système. Dans le cadre des propositions relatives, l'étude des relateurs est aussi un moyen de déceler d'éventuelles régularités dans l'emploi des pronoms relatifs. En outre, le degré de complexité des phrases (cumul de propositions, enchâssements, etc.) doit être pris en compte si l'on veut faciliter leur traitement par les grammaires du système. C'est aussi lors de cette étude qu'il est possible de vérifier si le corpus présente ou non certains types de propositions (conjonctives, relatives, interrogatives indirectes, infinitives).

Dans un deuxième temps, cette étape morpho-syntaxique comprend l'étude du comportement des verbes. On identifie ici les temps, modes et voix utilisés dans le corpus ainsi que la complémentation des verbes. Les caractéristiques de com-

plémentation (syntaxiques et sémantiques) peuvent être comparées à celles indiquées dans les ressources lexicales et terminologiques pour identifier les lacunes ainsi que les conflits.

Une troisième phase est consacrée au syntagme nominal. En effet, l'étude terminologique étant basée sur des *patrons,* certains syntagmes ont déjà pu être identifiés. Il semble toutefois important de recenser les autres possibilités de structures des syntagmes longs y compris des phrases averbiales pour vérifier des récurrences possibles intéressantes si l'on envisage de créer une mémoire de traduction réutilisable.

Une attention particulière est accordée à l'étude des pronoms anaphoriques (autres que les relateurs) afin de détecter d'éventuelles régularités dans la position du référent pour faciliter le traitement des anaphores par le système. Une précédente étude de ce dernier point réalisée par AÉROSPATIALE sur des documents normatifs a révélé des régularités importantes en terme de position des référents (90 % des référents étaient en position dernier objet de la phrase précédente) montrant ainsi l'intérêt de l'étude de ces pronoms pour une résolution facilitée des anaphores.

Les observations morpho-syntaxiques qu'AÉROSPATIALE se propose de faire n'ont pas toutes le même degré de faisabilité ; ceci est dû à l'identification parfois difficile des phénomènes recherchés. Ce problème est, de surcroît, amplifié par les volumes considérés et l'automatisation de l'étude de corpus pose des problèmes qui ne sont pas encore résolus.

Étude de structures sémantico-syntaxiques particulières

Cette troisième partie de l'étude de corpus est motivée par les observations qui ont pu être faites lors d'évaluations et par des préoccupations n'ayant pas forcément de lien direct avec la TA.

L'évaluation de systèmes de TA a révélé le problème important du traitement des *énumérations.* L'identification des structures énumératives les plus courantes permettrait d'en améliorer le traitement (identification des introducteurs d'énumérations, factorisation du verbe ou non, etc.). Un complément à cette étude serait de permettre au système d'identifier le début et la fin des énumérations afin que celles-ci bénéficient d'un traitement spécifique intégrant les données obtenues par l'étude de leurs structures. Il semble, en effet, que le problème majeur posé par les énumérations soit celui de leur reconnaissance par le système qui amalgame alors entre eux, ou avec les phrases suivantes, des éléments énumérés.

Un autre problème identifié lors des évaluations est celui des *titres* de documents, de chapitres, de paragraphes, etc., ayant pour caractéristique l'absence de marqueur de fin de séquence. Il est donc intéressant de dépister les caractéristiques formelles ou linguistiques qui pourraient permettre au système d'identifier de façon fiable les titres pour éviter l'amalgame de celui-ci avec la séquence qui suit.

La troisième structure recherchée est celle des *définitions.* Ces dernières n'ont, *a priori,* aucune incidence sur le processus de traduction mais leur étude est réalisée

dans une perspective d'acquisition de connaissances à des fins documentaires et d'apprentissage terminologique. Il est, en effet, tout à fait intéressant d'extraire directement des corpus traités les informations susceptibles d'aider à l'enrichissement des terminologies ou d'un thésaurus société. Nous rapprocherons cette étude des définitions avec celle des segments entre parenthèses qui se révèlent eux aussi d'un intérêt lexical et sémantique non négligeable (développés de sigles, définitions, synonymes, équivalents dans une langue étrangère, etc.).

Nous voyons que l'ensemble des informations morpho-syntaxiques obtenues sur un corpus peuvent être répercutées de façon différente au niveau du système proprement dit. Elles peuvent impliquer des enrichissements lexicaux et terminologiques ; elles peuvent restreindre certaines règles grammaticales, les enrichir au moyen d'heuristiques (anaphores) ou de préférences (sous-catégorisation) ; elles peuvent aider à l'identification des séquences dont on sait qu'elles nécessitent un traitement particulier, etc.

En ce qui concerne l'application de cette méthode sur un corpus réel, nous avons déjà signalé qu'elle avait nécessité quelques remaniements et travaux supplémentaires pour permettre une automatisation partielle de l'étude. C'est cette mise en œuvre qui a motivé un quatrième volet appelé *formats* qui a engendré, par la suite, une réorganisation des trois parties de l'étude que nous venons d'expliciter.

Mise en œuvre et automatisation partielle de l'étude de corpus

Dans le cadre du projet EUREKA GRAAL, les partenaires français du consortium (EDF, GSI-ERLI et AÉROSPATIALE) ont décidé de rendre automatisables le plus d'aspects possibles de l'étude de corpus en raison des volumes traités. Lorsque l'on automatise un processus, il est impératif de se pencher sur le format des documents qui doivent être manipulés : l'option SGML a été choisie. Le quatrième volet *format* découle directement de ce choix d'automatisation. Nous ne détaillerons pas ce dernier volet qui consiste à prendre en compte les aspects de compatibilité des traitements de textes et machines utilisées lors de la saisie des documents, ceux de l'identification des parties non ASCII (schémas et images) et de la mise au format SGML des documents qui posent des problèmes non spécifiques à notre sujet.

Au delà des problèmes cruciaux des formats de données, l'automatisation de l'étude a nécessité la mise en œuvre de ressources lexicales et grammaticales exploitables automatiquement ainsi qu'une réorganisation de la méthode directement liée à l'utilisation de ces ressources et à l'usage du format SGML.

Il était, en effet, intéressant de profiter des avantages de SGML pour identifier, partiellement du moins, certaines des structures sémantico-syntaxiques envisagées dans la section précédente sur la base de caractères typiques (: et - pour les énumérations, () et [] pour les segments entre parenthèses, combinaisons numériques et/ou alphabétiques pour les titres, etc.). De plus, l'intervention de ressources grammaticales a permis une lemmatisation du corpus facilitant grandement les études lexicales et la comparaison avec les ressources lexicales existantes ainsi qu'une analyse systématique des composants de la phrase en parties du discours (verbe, préposition, nom, adjectif, conjonction, adverbe) permettant des explorations multiples sur la base

de requêtes simples. Les principales étapes des études de corpus réalisées en automatique sont donc les suivantes :

Mise au format SGML :
- normalisation des caractères ;
- découpage du texte en séquences d'analyse ;
- numérotation des séquences ;
- identification des titres, énumérations et légendes.

Analyse linguistique :
- lemmatisation et attribution d'une catégorie ;
- comparaison avec les ressources lexicales et production d'une liste des inconnus ;
- identification des syntagmes nominaux correspondant aux *patrons* définis.

Ces opérations permettent déjà d'obtenir des données statistiques sur le corpus étudié relatives au nombre de mots, au nombre de séquences, au nombre moyen de mots par séquence, aux fréquences lexicales, à la répartition catégorielle, au nombre d'énumérations, au nombre d'ambiguïtés catégorielles possibles, au nombre de sigles identifiés comme tels, aux flexions utilisées.

L'automatisation du processus d'étude de corpus donne donc lieu à la génération de plusieurs fichiers issus du corpus original (fichier des mots inconnus, fichier des catégories par phrase, etc.) à partir desquels on peut envisager des requêtes plus ou moins complexes permettant d'identifier les phénomènes décrits dans les sections sur la démarche méthodologique. La définition des critères d'identification est actuellement en cours et des difficultés sont à prévoir pour certains phénomènes (anaphores, ellipses, etc.).

Études de corpus : perspectives

L'étude de corpus, telle qu'elle est décrite dans cet article, n'est pas suffisante pour permettre à l'utilisateur de définir l'ensemble des contraintes et besoins linguistiques. Dans le cadre de la participation à l'élaboration d'une application de traduction automatique, une étude traductologique s'impose. Elle sera réalisée ultérieurement. Toutefois, les résultats que l'on peut espérer obtenir sur une étude de corpus monolingue constituent des données pertinentes : elles peuvent avoir une influence directe sur l'élaboration des grammaires d'un système de TA ainsi que sur l'organisation des ressources lexicales et terminologiques qui seront utilisées au niveau de l'analyse. L'étude de corpus monolingue est le moyen pour l'utilisateur d'identifier ses contraintes linguistiques (qualité linguistique du texte source), l'étude traductologique, le moyen d'identifier ses besoins (qualité de la traduction).

Dans le cadre de l'évaluation des systèmes de traduction automatiques, les études de corpus représentent une base de travail indispensable si l'on désire évaluer les capacités d'un système à traiter les textes de l'utilisateur et non pas sa capacité à traiter tout type de texte dans tout type de domaines. Seuls les travaux d'évaluation fondés sur les corpus permettront de définir les améliorations indispensables à l'adaptation des systèmes aux besoins des utilisateurs.

Références

CHEVALIER, M., ISABELLE, P., LABELLE, F. et C. LAINE (1981) : « La traductologie appliquée à la traduction automatique », *Meta*, vol. 26, n° 1, Montréal, Les Presses de l'Université de Montréal, pp. 35-47.

DYSON, M. C. et J. HANNAH (1987) : « Toward a Methodology for the Evaluation of Machine Assisted Translation Systems », *Computers and Translation*, vol. 2, pp. 163-176.

HUTCHINS, W. J. et H. L. SOMERS (1992) : *An Introduction to Machine Translation*, Londres, Academic Press.

KING, M. (dir) (1987) : *Machine Translation Today : The State of Art, Proceedings of the Fourth Lugano Tutorial*, Edimbourg, Edinburgh Information Technology Series, Edinburgh University Press.

KING, M. et K. FALKEDAL (1990) : « Using Test Suites in Evaluation of Machine Translation Systems », Hans Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, University of Helsinki, pp. 211-216.

KITTREDGE, R. et J. LEHRBERGER (1982) : *Sublanguages, Studies of Language in Restricted Semantic Domains*, Berlin, De Gruyter.

KOCOUREK, R. (1991) : *La langue française de la technique et de la science*, Wiesbaden, Brandstetter Verlag.

LOFFLER-LAURIAN, A. M. (1982) : « Traduction automatique de textes techniques et analyse d'erreurs », *Contrastes*, supplément A1, pp. 45-58.

LOFFLER-LAURIAN, A.M. (1983) : « Pour une typologie des erreurs dans la traduction automatique », *Multilingua*, vol. 2, n° 2, pp. 65-78.

NIRENBURG, S. (1987) : *Machine Translation, Theoretical and Methodological Issues*, Cambridge, Cambridge University Press.

PALMER, M. et T. FININ (1990) : « Workshop on the Evaluation of Natural Language Processing Systems », *Computational Linguistics*, n° 16, Sept. 1990, pp. 175-181.

3

Des grammaires réutilisables pour la TA – et d'autres applications de TAL

Dominique ESTIVAL

ISSCO, Université de Genève, Genève, Suisse

• *Abstract* •

I propose here to present one aspect of the European research project GRAAL (Grammars which are Reusable for the Automatic Analysis of Language). The aim of this project is to provide a linguistic toolbox, consisting in modules for Natural Language Processing that will serve to build various NLP applications of different types.

The GRAAL grammar writing formalism should allow industrial partners to benefit from recent results in NLP research. Methods of analysis and representation which have already proved useful in research systems can now be applied in real size industrial system where their utility and dependability can be further demonstrated.

The theoretical basis of the GRAAL formalism are those of "Typed Feature Structures" and of "Unification-based system", extended with constraints. The grammars can thus be 'declarative', 'reversible' and 'modular'. Grammar modularity is the key concept, which allows a "core" grammar, modified by extensions which are specific to an application or a set of applications.

One of the targeted types of applications is MT, and the systems which are foreseen will for now be concerned primarily with French, English and Italian.

Le projet Graal

Nous présentons ici un des aspects du projet de recherche européen EUREKA intitulé GRAAL, *Grammaires réutilisables pour l'analyse automatique du langage.* Le but de ce projet est de constituer une boîte à outils linguistiques, consistant en modules de traitement automatique du langage naturel, qui puissent servir à construire différentes applications et différents types d'application de TAL.

Ce projet, qui a débuté en septembre 1992, s'inscrit dans le programme de recherche européen EUREKA, qui encourage les interactions entre l'industrie et le milieu universitaire. C'est donc un programme de recherche appliquée, et le consortium GRAAL comprend à la fois :

- des partenaires industriels : AÉROSPATIALE (France); EDF (France) ; FIAT (Italie) ; GSI-ERLI (France) ; NOKIA (Finlande) ;
- et des partenaires académiques : LINGSOFT (Finlande) ; ISSCO (Suisse) ; IRIT (France) ; IRST (Italie) ; ILSP (Grèce) ; ILTEC (Portugal).

Comme pour tous les projets du programme EUREKA, l'objectif est de faire bénéficier le milieu industriel des résultats de recherches récentes et qui ont déjà abouti à la validation de concepts dans le cadre d'institutions universitaires ou académiques. Pour le projet GRAAL en particulier, en se basant sur les recherches récentes effectuées en TAL, on peut délimiter trois objectifs principaux :

1. fournir un formalisme grammatical pour construire des grammaires qui peuvent être réutilisées, avec un *minimum de travail supplémentaire* dans diverses applications de TAL ;
2. fournir aux constructeurs d'applications une *boîte à outils* pour écrire de telles grammaires ;
3. créer des grammaires noyaux qui seront utilisées avec des extensions différentes, dans différentes applications.

Au niveau européen, on encourage les objectifs de recherche qui permettent de recouper les travaux d'autres projets et de réutiliser les résultats obtenus dans des domaines voisins. Dans cette optique, GRAAL a des liens privilégiés avec les projets suivants :

- GENELEX (projet EUREKA) : lexiques génériques et réutilisables ;
- TRANSTERM (projet LRE) : outils généraux pour extraire, manipuler et gérer les données terminologiques.

Les applications envisagées dans la partie applicative du projet sont bien entendu définies selon les besoins des partenaires industriels. Elles recouvrent des domaines du TAL assez variés :

- TA (français <—> anglais) ;
- extraction des connaissances (italien, anglais) ;
- dialogue homme-machine pour l'interrogation de bases de données (italien) ;
- indexation de textes et gestion documentaire (français, anglais) ;
- indexation de textes et extraction des connaissances (finnois, anglais).

La TA n'est donc que l'une de ces applications, mais les grammaires écrites pour l'application de TA doivent posséder un noyau commun avec les grammaires pour les autres applications traitant des mêmes langues. Les systèmes envisagés concernent pour le moment principalement le français, l'anglais et l'italien, mais on prévoit aussi le développement de grammaires pour le grec, le portugais, le finnois, et ultérieurement, l'allemand.

Architecture de GRAAL

L'architecture du projet est modulaire, et se compose principalement de trois bancs de travail (*WorkBenches*) :

- WB1 : développement des grammaires ;
- WB2 : développement des applications ;
- WB3 : maintenance des applications.

L'organisation de chaque banc de travail est, elle aussi, modulaire, de façon à faciliter la réutilisation des ressources lexicales ou terminologiques, et des bases de données de connaissances du domaine spécifiques à chaque application. Nous ne nous intéresserons ici qu'au WB1, qui concerne le développement des grammaires, et nous présenterons le formalisme pour les grammaires GRAAL, en gardant à l'esprit que les connexions avec des lexiques et des bases de données préexistantes sont prévues et doivent être accessibles à l'utilisateur dès ce stade.

Formalisme GRAAL TFS

Le formalisme GRAAL pour l'écriture des grammaires doit permettre à des industriels de bénéficier des résultats de recherches récentes en TAL en appliquant des méthodes d'analyse et de représentation qui ont déjà fait leurs preuves dans des systèmes de recherche, à un niveau où l'on peut maintenant démontrer leur utilité et leur fiabilité pour des systèmes commerciaux de taille réelle (Alshawi 1992, Alshawi *et al.* 1991, Carpenter 1992b, Carroll *et al.* 1988, Grover *et al.* 1989, Shieber 1988).

Le formalisme GRAAL se fonde donc sur les bases théoriques des *structures de traits typés* (Carpenter 1992a, Emele et Zajac 1990, Polard 1992) et sur les systèmes basés sur l'*unification* (Johnson 1988, Pereira et Shieber 1984, Shieber 1986). Le formalisme TFS (*Typed Feature Structures*) est étendu par un ensemble (limité) de *contraintes*, permettant notamment d'exprimer les contraintes de précédence linéaire séparément des relations de dominance (Emele *et al.* 1992). Une des caractéristiques essentielles du formalisme TFS de GRAAL est l'utilisation de la *hiérarchie des types* pour les inférences.

Les grammaires sont ainsi *déclaratives, réversibles* et *modulaires.* Ces qualités sont en effet nécessaires pour assurer certaines fonctionnalités indispensables pour des grammaires qui doivent, par définition, fonctionner dans différents environnements et pouvoir être modifées de façon transparente au gré des applications les incorporant.

- déclarativité : séparer le contrôle de l'information et les données linguistiques ;
- réversibilité : la même grammaire ou le même module grammatical peut être utilisé indifféremment pour l'analyse et pour la synthèse ;
- modularité : des segments de la grammaire peuvent être délimités et remplacés, ou des extensions peuvent être ajoutées.

La modularité des grammaires est en fait le concept clé qui permet d'avoir une grammaire « noyau », modifiée par des extensions spécifiques à une application ou un ensemble d'applications.

Les grammaires GRAAL

Une grammaire GRAAL se compose de trois parties :

- la hiérarchie des types ;
- la spécification des traits appropriés pour chaque type ;
- les « g-rules » (règles proprement dites).

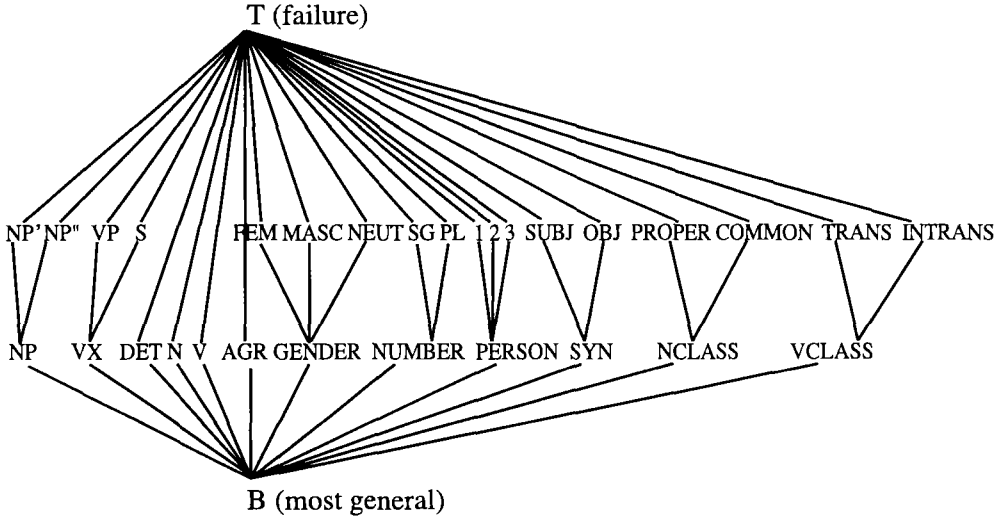
Le lexique peut être compilé et inséré dans la hiérarchie des types. Pour illustrer ces différentes parties, nous prendrons comme exemple une grammaire écrite dans le formalisme TFS de GRAAL, une petite grammaire complète qui permet de décrire (analyser et générer) les phrases suivantes :

(1) *The operator starts the engine.*
The operator stops the printer.
John starts the engine.
Mary stops the printer.

Hiérarchie des types

Une grammaire se compose d'éléments (les types) organisés en un treillis (la hiérarchie), que l'on peut représenter graphiquement comme en (2)[1] :

(2) Treillis de la hiérarchie



1. Cette hiérarchie peut être encodée de manière extrêmement efficace selon la procédure décrite par Aït-Kaçi *et al.* (1989).

Cette hiérarchie de types, qui correspond donc à la liste complète de tous les éléments de la grammaire et indique les relations entre ceux-ci, peut être décrite de manière informelle comme suit :

- Les types T (= TOP, échec) et B (= BOTTOM, toujours vrai) sont des types spéciaux, toujours présents dans n'importe quel système.
- Les types DET (déterminant), N (nom), V (verbe), NP (syntagme nominal), VP (syntagme verbal) et S (phrase), ainsi que NP', NP" et VX, représentent les catégories syntaxiques :
  - les types DET, N et V sont les catégories lexicales, les autres étant des catégories syntagmatiques.
  - les types DET, N, V, sont des types incompatibles.
  - le type VX est un supertype de VP et de S : les types VP et S héritent donc de certains des traits de VX.
  - le type NP a deux sous-types, NP' et NP", qui héritent de certains de ses traits.
- Les types AGR, GENDER, NUMBER, PERSON, SYN, SUBJ, OBJ, PROPER, etc., ne sont pas des catégories syntaxiques, mais des traits grammaticaux :
  - certains types (par exemple : GENDER, SYN) ont des sous-types, d'autres (comme : AGR, PROPER) n'en ont pas.
  - certains types (AGR, par exemple) ont des traits spécifiés, d'autres (comme : FEM) non.

Déclarations des types

Le grammairien qui écrit une grammaire énonce ces relations entre types qui définissent sa grammaire grâce à un ensemble de déclarations, soit pour notre exemple l'ensemble donné en (3).

Les types spéciaux T et B sont fournis par l'environnement GRAAL et automatiquement insérés dans la hiérarchie des types. Des types supplémentaires permettant de manipuler plus facilement certains objets (par exemple, LIST pour traiter des listes et STRING pour traiter des chaînes de caractères représentant les items lexicaux, voir section suivante) sont aussi prédéfinis pour l'utilisateur et fournis par l'environnement.

(3)

```
Subtypes   NP (NP' NP")
Subtypes   NP' (T)
Subtypes   NP" (T)
Subtypes   VX (S, VP)
Subtypes   VP (T)
Subtypes   S (T)
Subtypes   DET (T)
Subtypes   N (T)
Subtypes   V (T)
Subtypes   AGR (T)
Subtypes   GENDER (FEM, MASC, NEUT)
Subtypes   FEM (T)
```

```
Subtypes  MASC (T)
Subtypes  NEUT (T)
Subtypes  NUMBER (SG, PL)
Subtypes  SG (T)
Subtypes  PL (T)
Subtypes  PERSON (1, 2, 3)
Subtypes  1 (T)
Subtypes  2 (T)
Subtypes  3 (T)
Subtypes  SYN (SUBJ, OBJ)
Subtypes  SUBJ (T)
Subtypes  OBJ (T)
Subtypes  NCLASS (PROPER, COMMON)
Subtypes  PROPER (T)
Subtypes  COMMON (T)
Subtypes  VCLASS (TRANS, INTRANS)
Subtypes  TRANS (T)
Subtypes  INTRANS (T)
```

Spécification des traits appropriés : FAS

Quand un type porte des traits, ceux-ci peuvent être spécifiés grâce aux déclarations FAS (*Feature Appropriateness Specification*). Pour notre grammaire-exemple, les traits à spécifier sont les suivants :

(4)

```
Features NP    ( (constituents LIST) (class NCLASS)
                 (agreement AGR) (syn SYN) )
Features N     ( (string STRING) (class NCLASS)
                 (agreement AGR) )
Features VX    ( (constituents LIST) (class VCLASS)
                 (agreement AGR) )
Features V     ( (string STRING) (class NCLASS)
                 (agreement AGR) )
Features DET   ( (string STRING) (agreement AGR) )
Features AGR   ( (gen GENDER) (num NUMBER)
                 (per PERSON) )
```

Ces traits spécifient la liste des constituants (*constituents*) pour les catégories syntagmatiques et la chaîne de caractères de l'entrée lexicale (*string*) pour les catégories lexicales[2]. Ils précisent aussi les spécifications de traits d'accord (*agreement*) et la classe syntaxique (*class*) pour certaines catégories (avec pour valeur le type NCLASS ou VCLASS), ainsi que la fonction syntaxique (*syn*) pour les syntagmes nominaux.

Les traits spécifiés pour un supertype n'ont pas besoin de l'être pour ses sous-

2. Par définition seuls les types lexicaux peuvent porter le trait *string*, pour lequel le type STRING est prédéfini comme étant approprié.

types. Par exemple, les types NP', NP", S et VP héritent automatiquement des traits spécifiés pour leur supertype (NP et VX). On a donc, implicitement :

(5)

```
Features NP'    ( (constituents LIST) (class NCLASS)
                  (agreement AGR) (syn SYN) )
Features NP"    ( (constituents LIST) (class NCLASS)
                  (agreement AGR) (syn SYN) )
Features S      ( (constituents LIST) (class VCLASS)
                  (agreement AGR) )
Features VP     ( (constituents LIST) (class VCLASS)
                  (agreement AGR) )
```

Même avec un si petit exemple, on voit bien qu'il devient rapidement très difficile à un utilisateur de s'assurer que toutes ces déclarations et spécifications sont compatibles entre elles et n'entraînent pas de conflits. C'est pourquoi, outre les outils nécessaires à l'édition des différentes parties de la grammaire, l'environnement du banc de travail contient des outils de vérification qui assurent la cohérence de l'ensemble.

Les règles de grammaires : G-RULES

La troisième partie d'une grammaire GRAAL contient ce que l'utilisateur considérera sans doute comme les règles proprement dites. Cependant, une règle de grammaire n'est ici à strictement parler qu'une spécification plus poussée de certains types, ceux qui représentent des catégories syntagmatiques.

Les *g-rules* peuvent contenir, d'une part, des indications qui ne figurent pas dans les déclarations FAS, forçant ainsi la cooccurrence de certains traits dans certaines configurations (par exemple, l'accord sujet-verbe dans la règle (6)) et, d'autre part, des contraintes supplémentaires qui ne sont pas exprimables immédiatement dans un système TSF pur.

Par exemple, les relations de dominance peuvent être encodées de manière classique grâce à un trait tel que *constituents* pour lequel, comme nous l'avons vu, le type LIST est approprié, et qui peut donc représenter la liste des constituants d'un syntagme.

L'ordre linéaire pourrait aussi être encodé dans la définition du type lui-même : il suffirait de définir un type (par exemple, LIST') comme étant une liste ordonnée, mais cela ne présenterait pas la flexibilité de description que l'on a quand les deux types de relation, dominance et précédence, sont séparés. Comme, en effet, il s'avère toujours plus commode de pouvoir séparer, dans une représentation linguistique, les phénomènes d'ordre linéaire de ceux de dominance immédiate (voir GPSG, Gazdar *et al.* 1985 ; HPSG, Pollard et Sag 1987 ; LFG, Bresnan et Kaplan 1982 ; ainsi que les différents principes et paramètres de GB), nous proposons comme contrainte prédéfinie celle qui permet de représenter les relations de précédence sous la forme {x@ < y@}. Cette contrainte peut ne pas être présente, comme dans (7) qui ne comporte qu'un seul constituant, ou pourrait être omise si la langue décrite par la grammaire a un ordre des mots plus ou moins libre.

(6) Phrase

```
S (constituents => [x@NP (agreement => z@AGR (gen => v@
                                                num => w@)
                          syn => SUBJ )
                    y@VP (agreement -> z@AGR)]
agreement => z@AGR)
{x@ < y@}
```

(7) NP ne contenant qu'un nom propre

```
NP'( constituents => [x@N (agreement => z@AGR,
                           class => PROPER)]
     agreement => z@AGR)
```

(8) NP contenant un déterminant et un nom commun

```
NP"( constituents => [x@N (agreement => z@AGR
                                        class => COMMON)
                      y@DET (agreement => z@AGR) ]
     agreement => z@AGR)
  {y@ < x@}
```

(9) Simple VP contenant un verbe suivi de son objet direct

```
VP ( constituents => [x@V (agreement => z@AGR
                                      class => TRANS)
                      y@NP (syn => OBJ) ]
     agreement => z@AGR)
  {x@ < y@}
```

Entrées lexicales

De la même manière, les entrées lexicales ne sont en fait que des spécifications plus précises de certains types, ceux qui possèdent le trait *string*. On voit dans les exemples donnés en (10) que les traits qui sont spécifiés dans les FAS comme étant appropriés pour ces types peuvent recevoir ici une valeur particulière (par exemple, les traits de genre, nombre et personne pour les noms propres), ou être laissés sous-déterminés (comme les traits d'accord pour l'article en anglais, ou le trait de genre pour le nom *operator*).

(10)

```
N (string => John
   class => PROPER
   agreement => (gen => MASC, per => 3, num => SG) )

N (string => Mary
   class => PROPER
   agreement => (gen => FEM, per => 3, num => SG) )

N (string => operator
   class => COMMON
   agreement => (per => 3, num => SG) )
```

```
N (string => engine
   class => COMMON
   agreement => (gen => NEUT, per => 3, num => SG) )

N (string => printer
   class => COMMON
   agreement => (gen => NEUT, per => 3, num => SG) )

V (string => starts
   class => TRANS
   agreement => (per => 3, num => SG) )

V (string => stops
   class => TRANS
   agreement => (per => 3, num => SG) )

DET (string => the)
```

Modularisation de la grammaire

Puisque le but du projet GRAAL est que l'utilisateur puisse assez facilement, d'une part, réutiliser certaines parties d'une grammaire et, d'autre part, y ajouter des extensions, une grammaire GRAAL doit être modulaire et il est nécessaire de définir une manière assez simple d'ajouter des règles ou des ensembles de règles.

Pour ajouter une nouvelle règle de grammaire (au sens linguistique, c'est-à-dire pour traiter un phénomène) dans le formalisme que nous venons de définir, il faudra au moins :

- ajouter (au moins un) type pour chaque nouvelle règle ;
- spécifier les traits qui sont appropriés pour ce type ;
- ajouter la g-rule le spécifiant ;
- ajouter si nécessaire les entrées lexicales qui y font appel.

Nous donnerons ici comme exemple les étapes requises pour que notre petite grammaire puisse maintenant décrire des phrases contenant des verbes intransitifs comme dans (11) et des constructions à double objet comme dans (12).

(11) *The engine fails.*
The printer stops.

(12) *Mary gives the operator a printer.*
The operator gives John the engine.

Nouvelles déclarations de types

(13) Subtypes SYN (SUBJ, OBJ, IOBJ)
Subtypes SUBJ (T)

```
Subtypes  IOBJ (T)
Subtypes  VCLASS (TRANS, INTRANS, DTRANS)
Subtypes  TRANS (T)
Subtypes  INTRANS (T)
Subtypes  DTRANS (T)
Subtypes  VP (VP1, VP2, VP3)
Subtypes  VP1 (T)
Subtypes  VP2 (T)
Subtypes  VP3 (T)
```

Nouvelles G-RULEs

(14) VP contenant un verbe intransitif

```
VP1( constituents => [x@V (agreement => z@AGR
                                        class => INTRANS)]
     agreement => z@AGR)
```

(15) VP contenant un verbe suivi de son objet direct

```
VP2( constituents => [x@V (agreement => z@AGR
                                        class => TRANS)
                          y@NP (syn => OBJ) ]
     agreement => z@AGR)
   {x@ < y@}
```

(16) VP contenant un verbe suivi de ses objets (indirect et direct)

```
VP3( constituents => [x@V (agreement => z@AGR
                                        class => DTRANS)
                          y@NP (syn => IOBJ)
                          w@NP (syn => OBJ) ]
     agreement => z@AGR)
   {x@ < y@, y@ < w@}
```

Nouvelles entrées lexicales

(17)

```
V (string => gives
   class => DTRANS
   agreement => (per => 3, num => SG) )

V (string => stops
   class => INTRANS
   agreement => (per => 3, num => SG) )

V (string => fails
   class => INTRANS
   agreement => (per => 3, num => SG) )

DET (string => a)
```

Conclusion

Les résultats obtenus aujourd'hui portent principalement sur trois plans. Cet article a décrit le premier volet, la définition du formalisme grammatical GRAAL TFS.

Les spécifications de la *boîte à outils* GRAAL constituent le deuxième volet et, ainsi qu'il a été dit au début, cette boîte est divisée en trois bancs de travail : pour le développement des grammaires, pour le développement des applications et pour la maintenance des applications. Un prototype pour le premier banc de travail est en cours de développement, et les outils nécessaires au développement des grammaires, en particulier ceux qui concernent les vérifications de cohérence, sont presque complètement spécifiés.

Les prototypes pour les deux autres bancs de travail sont eux aussi à l'étude, et incluront certains des outils du premier.

Comme un certain nombre d'applications prévues reposeront sur des bases de textes existantes et que les grammaires d'application devront contenir des extensions qui leur seront spécifiques, le projet requiert une analyse de corpus. Ce troisième volet est déjà bien avancé (voir l'article d'Éva Dauphin dans cet ouvrage).

Le travail doit maintenant se poursuivre et concerne la mise en œuvre des moteurs linguistiques (analyseur, générateur et module de transfert pour la TA) ; la mise en œuvre des outils dans la *boîte à outils* et, bien sûr, la spécification et la mise en œuvre des applications.

De plus, c'est surtout la rédaction des grammaires noyaux avec le formalisme, tel qu'il a été défini et présenté ici, qui va nous occuper dans l'avenir immédiat, avec la perspective de les compléter par des extensions dans un avenir proche.

Références

AÏT-KAÇI, H., R. BOYER, P. LINCOLN et R. NASR (1989) : « Efficient Implementation of Lattice Operations », *ACM Transactions on Programming Languages and Systems*, vol. 11, n° 1.

ALSHAWI, H. (Ed.) (1992) : *The Core Language Engine*, ACL-MIT Press Series in Natural Language Processing.

ALSHAWI, H., ARNOLD, D. J., BACKOFEN, R., CARTER, D. M., LINDOP, J., NETTER, K., PULMAN, S., TSUJII, J. et H. USZKOREIT (1991) : *EUROTRA ET6/1: Rule Formalism and Virtual Machine Design Study*, Final Report, CEC.

BOGURAEV, B., CARROLL, J., BRISCOE, T. et C. GROVER (1988) : « Software Support for Practical Grammar Development », *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, Budapest, pp. 54-57.

CALDER, J., KLEIN, E. et H. ZEEVAT (1988) : « Unification Categorial Grammar – A Concise, Extendable Grammar for Natural Language Processing », *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, Budapest, pp. 83-86.

CARPENTER, B. (1992b) : ALE : *The Attribute Logic Engine User's Guide*, Laboratory for Computational Linguistics, Carnegie Mellon University, Pittsburgh.

CARPENTER, B. (1992a) : *The Logic of Typed Feature Structures. With Applications to Unification Grammars, Logic Programs and Constraint Resolution*, Cambridge University Press.

CARROLL, J, BOGURAEV, B., GROVER, C. et T. BRISCOE (1988) : *A Development Environment for Large Natural Language Grammars*, Technical Report 127, Computer Laboratory, University of Cambridge.

EMELE, M., HEID, U., MOMMA, S. et R. ZAJAC (1992) : « Interactions Between Linguistic Constraints: Procedural vs Declarative Approaches », *Machine Translation*, 7, pp. 61-98.

EMELE, M. et R. ZAJAC (1990) : « Typed Unification Grammars », Hans Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, Helsinki.

GAZDAR, G., KLEIN, E., PULLUM, G. et I. A. SAG (1985) : *Generalized Phrase Structure Grammar*, Cambridge MA, Harvard University Press.

GERDEMAN, D. G. (1991) : *Parsing and Generation of Unification Grammars*, Cognitive Science Technical Report CS-91-06, University of Illinois at Urbana Champaign.

GROVER, C., BRISCOE, T., CARROLL, J. et B. BOGURAEV (1989) : *The Alvey Natural Language Tools Grammar (Second Release)*, Technical Report 162, Computer Laboratory, University of Cambridge.

JOHNSON, M. (1988) : *Attribute-value Logic and the Theory of Grammar*, CSLI Lecture Notes, 16, CSLI, Stanford California.

KAPLAN, R. et J. BRESNAN (1982) : « Lexical-functional Grammar: A Formal System for Grammatical Representation », J. Bresnan (dir), *The Mental Representation of Grammatical Relations*, Cambridge, MIT Press, pp. 173-281.

PEREIRA, F. et S. SHIEBER (1984) : « The Semantics of Grammar Formalisms Seen as Computer Languages », *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84)*, Stanford, pp. 123-129.

POLLARD, C. J. (1992) : « Sorts in Unification-based Grammar and What They Mean », Pinkal, M. et B. Gregor (dir), *Unification in Natural Language Analysis*, MIT Press.

POLLARD, C. et I. A. SAG (1987) : *Information-based Syntax and Semantics, vol. 1: Fundamentals*, CSLI Lecture Notes, 13, Stanford, CSLI and Chicago, University of Chicago Press.

SHIEBER, S. (1986) : *An Introduction to Unification-based Approaches to Grammar*, CSLI Lecture Notes, 4, CSLI, Stanford, California.

SHIEBER, S. (1988) : *CL-PATR User's Manual, Version 1.0.*, Artificial Intelligence Center and CSLI, Stanford California.

4

La génération multilingue : des convergences aux divergences

Liesbeth DEGAND

GMD/ Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Allemagne et UCL, Faculté de psychologie, Louvain-la-Neuve, Belgique

• *Abstract* •

In this paper we present some early results of our ongoing research on multilingual text generation. The multilingual architecture is based on functional typology which permits the sharing of resources on all levels of the linguistic system (lexicogrammar, semantics, conceptual base, ...) and also allows the cumulative addition of other languages without revising the overall system. We will then concentrate on the linguistic resources responsible for ensuring that texts are texts and not mere sequences of sentences. This "textuality" is an important issue to deal with when multilinguality is considered and raises several important questions. We must guarantee, first, that language-specific textuality is achieved for each language covered, and second, that the same text is generated in each language. We present some examples of texts from French, Dutch and German for which we provide an account of cross-linguistic convergences and divergences, primarily in the textual metafunction. We will give a motivation of the textual divergences encountered and suggest how they can be controlled in order to express them on the lexicogrammatical level in each of the language components.

Introduction

La génération multilingue est de plus en plus considérée par de nombreux chercheurs comme une alternative sérieuse à la traduction automatique. En effet, nous vivons dans un monde de plus en plus plurilingue (pensons à la Communauté européenne, par exemple), où de nouveaux textes sont créés en permanence et doivent être accessibles à des locuteurs de langues très diverses. Dans un tel contexte, la traduction devient très rapidement impraticable. Aussi, de nombreuses organisations ont-elles

décidé de passer de la traduction à la *production de textes*, procédé par lequel chaque texte est directement écrit dans chacune des langues visées.

À ce stade, une première question peut être posée : dans le contexte de la génération multilingue, comment faire face au nombre croissant de langues de travail ? (La même question se pose d'ailleurs pour les systèmes de traduction automatique.)

La génération multilingue s'est principalement concentrée sur le développement de systèmes capables de produire des textes dans plusieurs langues, et ce à partir d'une même base de connaissances. Peu de travail a été effectué jusqu'à ce jour pour rendre ces systèmes facilement extensibles à de nouvelles composantes linguistiques. En effet, la plupart des systèmes ont en fait une structure monolingue : ils ne constituent, en fait, qu'un assemblage de différentes composantes linguistiques sans aucun partage de ressources si ce n'est au niveau conceptuel (et parfois sémantique). Cependant, en se basant sur une théorie typologique adéquate du langage, il est possible de construire un système où les différentes langues peuvent à la fois converger et diverger en termes de base de connaissances, de sémantique, de grammaire, etc. L'approche que nous proposons, repose sur la notion linguistique de similarité et différence fonctionnelles : en saisissant en premier lieu les fonctions qu'accomplissent les langues (naturelles), nous atteignons un niveau de description linguistique qui s'étend à travers les langues de manière beaucoup plus efficace et plus globale que les descriptions à base structurelle. Nous montrerons également que cette architecture à typologie fonctionnelle permet l'addition cumulative de nouvelles composantes linguistiques sans révision globale du système de départ (et en réutilisant bon nombre de ressources déjà présentes dans le système).

La seconde question qui se pose dans le cadre de la génération multilingue est d'ordre textuel. Comment garantir que les *mêmes* textes sont générés dans les différentes langues ?

Il va sans dire que la notion d'*équivalence* des textes est cruciale. Il faut s'assurer que les textes produits dans les différentes langues puissent être reconnus comme étant le *même texte*, au-delà de la variation langagière. Une façon d'atteindre ce but est de définir des standards auxquels chacun des textes devrait répondre. Nous avons néanmoins opté pour une solution purement linguistique. Nous définissons différents niveaux de représentation linguistique pour chacun desquels équivalence et non-équivalence peuvent être mis en discussion, et nous générons les textes en fonction de ces critères. En effet, en représentant de manière explicite les similarités et différences entre les langues, et ce à chaque niveau de description, il devient plus aisé de clarifier de manière plus générale la notion d'équivalence, et dès lors d'élaborer des modèles plus adéquats du processus de traduction.

Le présent article est organisé de la manière suivante :

Après une brève description de notre système multilingue proprement dit, nous nous concentrons sur le niveau de la lexicogrammaire. Nous montrerons quelles sont les implications du choix d'une théorie fonctionnelle sur l'organisation générale de la grammaire. Nous décrivons ensuite de quelle façon une nouvelle langue peut être ajoutée au système, montrant ainsi que tout en partageant un certain nombre de ressources, chacune des langues garde toute sa spécificité.

Nous abordons ensuite les aspects plus spécifiquement textuels de la génération multilingue montrant que les principes de multilingualité énoncés au niveau de la lexicogrammaire peuvent et doivent être étendus au niveau de la sémantique (discursive). À l'aide d'exemples, nous nous attacherons plus spécialement à la description des divergences et des convergences textuelles qui peuvent exister entre des textes (originaux) en français et en néerlandais. Cette textualité (souvent divergente) est réalisée par un grand nombre de structures linguistiques au niveau de la lexicogrammaire. La question qui se pose alors est de savoir comment contrôler ces réalisations. En d'autres mots, comment passer du niveau relativement abstrait de la base textuelle aux spécifications lexicogrammaticales détaillées, et ce pour chacune des langues. Nous pensons que les propositions les plus détaillées qui ont été faites dans ce domaine jusqu'à présent (par exemple, Bateman *et al.* 1991a ; Hovy *et al.* 1992) n'offrent pas suffisamment de contrôle. Nous proposons dès lors une extension de ces systèmes basée sur le travail linguistique de Martin (1992). Finalement nous discutons les résultats de cette première ébauche de conception d'une sémantique discursive pour la génération multilingue.

Un système de génération à motivation fonctionnelle

Le système de génération multilingue que nous présentons ci-après est le résultat d'une collaboration entre GMD/IPSI (Darmstadt, Allemagne) et l'Université de Sydney. Il repose sur un système de génération de textes bien établi pour l'anglais, le système PENMAN (en développement à USC/ISI depuis 1980), qui a lui-même servi de point de départ au système KOMET pour l'allemand (développement en cours depuis 1989 à GMD/IPSI) et le projet multilingue à Sydney, qui a débuté en 1990 pour la génération de l'anglais, du japonais et du chinois. La théorie linguistique sous-jacente aux trois systèmes est la linguistique systémique fonctionnelle (Halliday 1978 ; 1985). Il s'agit d'une théorie multidimensionnelle en ce qu'elle est stratificationnelle et fonctionnelle.

Dans notre architecture, le système linguistique est composé de trois strates :

- *l'environnement sémantique* contient des unités de sens abstraites qui constituent l'environnement sémantique pour tout texte, et les principaux moyens pour le relier au contexte ;
- *l'interface sémantique* contient les ressources pour mettre en relation ces unités abstraites de sens avec la grammaire ;
- *la lexicogrammaire* contient les ressources grammaticales pour exprimer ces unités de sens en structures lexicales et syntaxiques.

Il s'agit donc d'une théorie linguistique générale qui prend en compte tous les niveaux de sens apparaissant dans la langue. Ceci est également reflété dans l'organisation fonctionnelle du modèle. En effet, les éléments de chacune des strates peuvent être décomposés selon trois domaines généralisés de sens : *les métafonctions idéationnelle, interpersonnelle et textuelle. Grosso modo,* la métafonction idéationnelle encode le contenu propositionnel de la langue, la métafonction interpersonnelle les attitudes et comportements des locuteurs, et la métafonction textuelle les aspects textuels de la communication (pour plus de détails sur la théorie des métafonctions, voir, par exemple, Halliday 1978). À chacune des strates, ces trois métafonctions ont des manifestations grammaticales différentes.

Au niveau de l'environnement sémantique (la strate la plus élevée), ces métafonctions sont reliées à trois bases fonctionnellement distinctes : la base *idéationnelle,* comparable à la base de connaissance, *interactionnelle*, comparable au modèle du locuteur *(user model)* et *textuelle*, comparable au modèle discursif *(discourse model).* Le système linguistique est donc organisé selon deux dimensions : la *profondeur* stratificationnelle et la *largeur* métafonctionnelle ; ceci est illustré dans la figure 1.

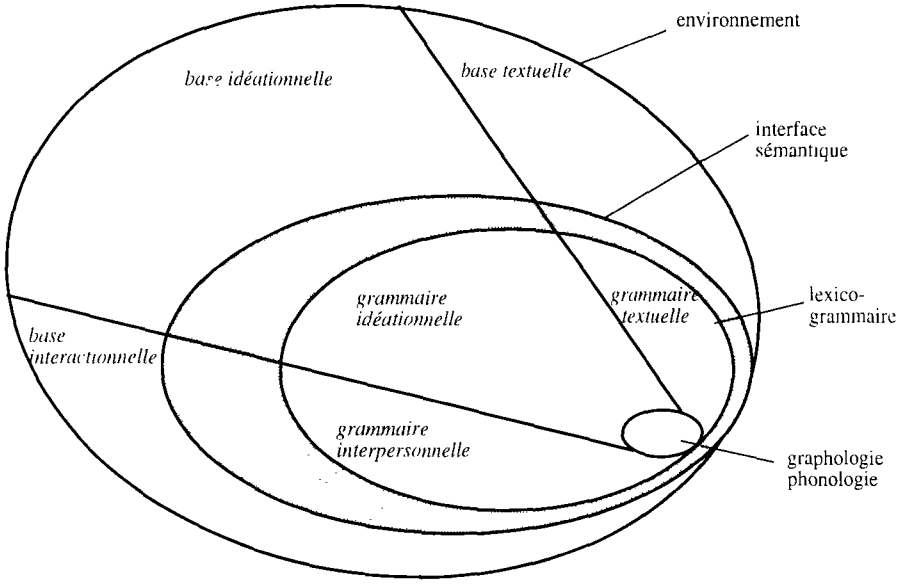


FIGURE 1 : Stratification des ressources linguistiques.

L'hypothèse théorique sous-tendant le système multilingue est que les similarités entre les langues sont fonctionnelles en premier lieu. Il est alors possible de partager cette fonctionnalité commune dans une large mesure, à tous les niveaux du modèle linguistique, alors que les réalisations structurelles peuvent très bien diverger. Ce fondement fonctionnel doit donc permettre de concevoir un système multilingue où les différentes langues peuvent aussi bien converger que diverger le long des différentes dimensions du système linguistique. La multilingualité de l'architecture repose donc sur le fait qu'elle supporte un partage maximal d'informations à travers les différentes langues tout en permettant aux descriptions individuelles d'être aussi divergentes et spécifiques qu'il s'avère nécessaire. En outre, cette architecture à typologie fonctionnelle permet l'addition cumulative d'autres composantes linguistiques sans révision du système de base et en réutilisant de nombreuses ressources déjà présentes dans le système ; nous reviendrons sur ce point dans la section suivante.

Convergences et divergences au niveau lexicogrammatical

Dans un premier temps, nous nous concentrerons sur les convergences et les divergences présentes entre les langues au niveau lexicogrammatical. Selon la théorie systémique, la strate lexicogrammaticale est composée de différents *rangs* (*ranks*). L'anglais, l'allemand

et le néerlandais, par exemple, possèdent quatre rangs : l'énoncé, le syntagme, le mot et le morphème. Les rangs sont organisés selon une hiérarchie décroissante. Les fonctions des unités d'un rang donné sont réalisées par les unités du rang inférieur c'est-à-dire, les fonctions au rang de l'énoncé sont réalisées par les unités appartenant au rang du syntagme, les syntagmes par les mots, et les mots par les morphèmes. Chacun de ces niveaux est organisé en une arborescence composée de *systèmes*. Un système est formé d'une disjonction d'au moins deux traits et des conditions sous lesquelles ces traits sont accessibles. Les sélections de ces traits sont réalisées sous forme de structures ou d'items lexicaux. Un système peut donc être considéré comme un expert dans un domaine donné, chaque système faisant un choix déterminant et contraignant la structure linguistique finale. Ainsi en traversant le réseau (c'est-à-dire, en faisant des choix), on accumule petit à petit l'information nécessaire à la forme finale. Nous verrons comment ces réseaux peuvent être semblables ou différents pour les différentes langues traitées par le système multilingue. Pour ce faire, nous décrirons comment nous avons ajouté une nouvelle composante linguistique, plus précisément un fragment de grammaire néerlandaise, au système multilingue tirant profit des ressources déjà présentes dans l'architecture.

Le point de départ de la grammaire multilingue est la grammaire anglaise Nigel (Matthiessen 1992). Il s'agit de la composante grammaticale partagée par toutes les autres composantes linguistiques du système. Des divergences par rapport à la grammaire anglaise sont indiquées par un label langagier, ce qui permettra une traversée du réseau grammatical spécifique à une langue. Lorsqu'aucune indication n'est donnée, le système est valable pour toutes les langues du système. L'extrait de réseau présenté dans la figure 2 montre comment un réseau de systèmes peut à la fois faire apparaître des convergences entre des langues typologiquement très différentes comme l'anglais, le japonais et le chinois, mais laisse aussi la possibilité de représenter les spécificités de chacune des langues.

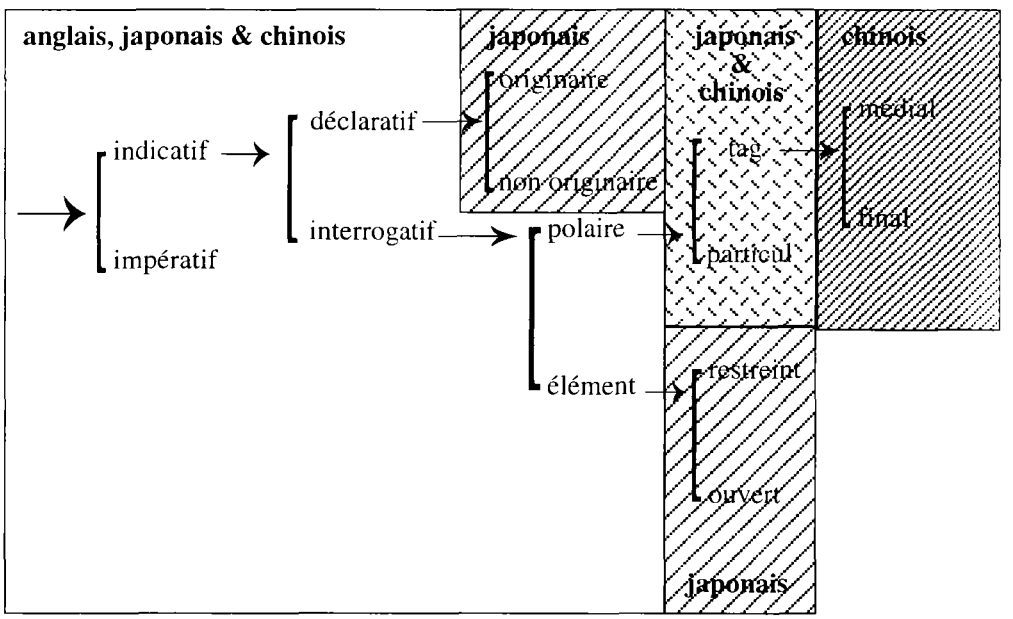


FIGURE 2 : Réseau multilingue de Mode pour l'anglais, le japonais et le chinois.

Il en va bien sûr de même pour des langues typologiquement proches. Afin d'illustrer que le système supporte effectivement l'addition cumulative de nouvelles langues sans révision globale, et analyser les implications que cette addition pourrait néanmoins avoir sur le système, nous allons décrire maintenant comment le système multilingue a été étendu pour inclure un fragment de grammaire néerlandaise (toujours en développement !). Ce travail a été effectué lors d'une expérience pendant laquelle un même texte a été généré en anglais, en allemand et en néerlandais (pour plus de détails, voir Degand 1993).

Une nouvelle composante linguistique pour le système multilingue

Les convergences entre les grammaires anglaise et néerlandaise se manifestent dans le système multilingue par l'absence de label langagier au niveau des systèmes ; ces systèmes sont valables pour les deux langues. Les différences entre les deux langues sont exprimées au niveau de la lexicogrammaire par des traversées différentes du réseau grammatical. En termes de systèmes, les modifications à apporter étaient généralement de quatre types. Il fallait :

- *couper des systèmes,* les systèmes grammaticaux décrivent des phénomènes qui ne sont pas pertinents pour la langue ajoutée. En termes systémiques, on dira que le réseau de la nouvelle langue présente moins de « finesse » (*delicacy*), c'est-à-dire, qu'il est moins détaillé en profondeur à cet endroit. Le néerlandais ne connaît pas, par exemple, le phénomène du *tagging*. Les systèmes décrivant des choix dans ce domaine ne pourront donc être traversés lors du traitement du néerlandais, mais pour l'anglais ces choix sont pertinents. C'est pourquoi ces systèmes seront marqués comme appartenant à la grammaire de l'anglais uniquement.

- *ajouter des systèmes,* c'est-à-dire faire des descriptions plus « délicates ». Le néerlandais établit une distinction entre le genre grammatical neutre et commun pour les noms, alors que l'anglais n'a pas de distinction de genres. Il faudra donc ajouter un tel système pour le néerlandais.

- *changer les manifestations structurelles (realization statements).* Ceci est le cas, lorsque les langues présentent des similarités fonctionnelles, mais les réalisations structurelles sont divergentes. Le passé perfectif existe en anglais comme en néerlandais, mais structurellement l'ordre de l'élément verbal non fini sera différent.

- *changer les conditions d'entrée.* Parfois les langues semblent faire les mêmes distinctions paradigmatiques (les systèmes sont pertinents pour les deux langues), mais elles diffèrent néanmoins dans le sens où les conditions d'entrée pour un « même » système ne sont pas équivalentes ; en d'autres mots, le contexte paradigmatique est différent.

Très souvent, les différences entre les langues apparaîtront comme une combinaison de ces quatre types de divergences.

Outre la confirmation que l'architecture à typologie fonctionnelle permet effectivement une extension aisée du système multilingue à de nouvelles langues, notre

expérience a également permis de vérifier trois principes concernant les similarités fonctionnelles, principes qui avaient été annoncés dans Bateman *et al.* (1991b).

1. Pour une strate donnée, il est plus probable que l'on rencontre ou que l'on doive faire apparaître des similarités fonctionnelles sur l'axe paradigmatique (en termes systémiques), plutôt que sur l'axe syntagmatique (en termes d'organisation structurelle). L'ordre structurel des mots dans une phrase est ici un exemple évident.

2. Pour une strate donnée, il est plus probable que l'on rencontre ou que l'on doive faire apparaître des similarités fonctionnelles au degré de « finesse » le plus bas, plutôt qu'au niveau le plus élevé. Toutes les langues inclues dans le système multilingue jusqu'à présent semblent faire une distinction entre le mode indicatif et le mode impératif. Mais en anglais, par exemple, on distingue entre différentes formes de l'impératif, alors que le néerlandais ne connaît qu'une seule forme.

3. Pour une strate donnée, il est plus probable que l'on rencontre ou que l'on doive faire apparaître des similarités fonctionnelles au rang le plus élevé plutôt qu'au rang le plus bas. Pour la grammaire, cela signifie que les différentes langues auront vraisemblablement plus de systèmes en commun appartenant au rang décrivant la proposition et que ce partage des ressources ira en diminuant au fur et à mesure que l'on descendra aux systèmes décrivant les groupes, les mots et les morphèmes.

Un dernier point que nous aimerions signaler à propos du développement de nouvelles composantes linguistiques dans ce cadre multilingue est que cette démarche apporte un gain de temps considérable pour le développement. Comme la fonctionnalité semble en effet être un élément plutôt commun entre les langues, de nouvelles composantes peuvent tirer un bénéfice substantiel du travail et de la recherche effectués pour les langues dont le développement est déjà en cours. Pour l'addition d'une langue proche au niveau typologique, telle que le néerlandais par rapport à l'anglais et l'allemand, nous estimons que le temps nécessaire pour atteindre un niveau de « couverture grammaticale » comparable à celui de la grammaire Nigel[1], qui est une des plus grandes grammaires computationnelles existant pour le moment, est de 1 personne/5 mois. En outre, ce temps de développement pourra encore être raccourci, lorsque d'autres composantes linguistiques seront ajoutées au système. En effet, il deviendra alors plus probable que l'on trouve, pour chaque langue ajoutée, une solution à un problème spécifique dans une des langues déjà développées. Le développement de nouvelles langues dans le système multilingue peut également être fort bénéfique aux langues qui se trouvent dans un stade de développement déjà plus avancé. Les « corrections » ayant principalement un caractère fonctionnel, la motivation fonctionnelle des décisions grammaticales se trouve améliorée pour le système dans son ensemble.

1. Une mesure du degré de développement d'une nouvelle composante grammaticale par rapport aux grammaires déjà existantes est fournie par le tableau des régions fonctionnelles de la grammaire (domaines dans la grammaire caractérisés par une unité sémantique et grammaticale), voir, par exemple, Matthiessen et Bateman (1991). Dans le stade actuel du développement de la grammaire néerlandaise, pratiquement toutes les régions fonctionnelles de la grammaire anglaise sont en partie couvertes, quelques régions répondant à certaines spécificités du néerlandais ont été ajoutées, alors que certaines régions spécifiquement anglaises ont été supprimées (Degand, en préparation).

La prochaine étape dans nos investigations multilingues a consisté à vérifier si les principes de multilingualité émis au niveau de la lexicogrammaire pouvaient être transposés à un niveau d'abstraction plus élevé, plus précisément au niveau de la sémantique discursive. Une première expérience dans ce sens portant sur le néerlandais, l'allemand et l'anglais (voir Bateman *et al.* 1993) semble le confirmer ; nous voulons ici pousser ces investigations plus loin et les appliquer à des langues typologiquement moins proches, le néerlandais et le français.

Convergences et divergences au niveau textuel

Afin d'être en mesure de générer des textes, et non de simples suites de phrases, nous nous sommes penchés sur les ressources linguistiques responsables de la textualité d'un texte. Les phrases qui forment ensemble un texte ont des propriétés collectives qui dépassent celles que l'on peut attribuer à chacune d'entre elles séparément. C'est un fait dont il faut tenir compte lorsque l'on travaille dans un cadre multilingue et il soulève plusieurs questions. Premièrement, comment s'assurer que l'on obtient une textualité satisfaisante dans chacune des langues traitées ? Les traductions directes très souvent ne s'arrêtent pas à ces aspects textuels, nous en donnerons des exemples plus loin. Deuxièmement, comment pouvons-nous garantir qu'un *même* texte est généré dans chacune des langues ?

Afin d'analyser ces questions, nous avons rassemblé un corpus contenant entre autres des textes donnant des informations générales sur les problèmes de santé. Ces textes sont produits en plusieurs langues européennes par l'Union européenne des Pharmacies sociales. Le genre des textes est didactique, donnant des informations médicales de base à des non-experts ; le registre (niveau de langage) est général et impersonnel. Les exemples que nous discuterons sont extraits d'un texte informatif sur les risques liés aux voyages, dans notre cas précis les risques liés à la chaleur :

Néerlandais

(n:1) *Blootstelling aan de zon kan verbrandingen veroorzaken.* (n:2a) *Een warm klimaat kan daarenboven ook ernstige problemen veroorzaken,* (n:2b) *zoals een hitteberoerte.* (n:3) *Om dit probleem te voorkomen moeten de lichamelijke activiteiten tijdens de heetste uren van de dag worden vermeden.* (n:4a) *Het is ook ten zeerste aangeraden om het hoofd af te dekken,* (n:4b) *veel water te drinken en* (n:4c) *wat zout in te nemen.* (n:5) *De wens om snel te bruinen kan u parten spelen.* (n:6) *Het is beter om u geleidelijk dag na dag aan de zon bloot te stellen.* (n:7a) *Men zal zich met een aan zijn huidtype en bruiningsgraad aangepaste crème insmeren en* (n:7b) *een zonnebril van goede kwaliteit dragen.*

(L'exposition au soleil peut provoquer des brûlures. Un climat chaud peut en outre également provoquer des problèmes graves comme le coup de chaleur. Pour prévenir ce problème, les activités physiques doivent être évitées pendant les heures les plus chaudes de la journée. Il est aussi vivement conseillé de se couvrir la tête, de boire beaucoup d'eau et de consommer du sel. Le désir de bronzer rapidement peut vous jouer de mauvais tours. Il vaut mieux vous exposer graduellement jour après jour au soleil. L'on s'enduira d'une crème adaptée à son type de peau et son degré de bronzage et (l'on) portera des lunettes solaires de bonne qualité.)

Français

(f:1a) Outre de légères brûlures, un climat chaud et l'exposition au soleil peuvent provoquer des troubles sérieux (f:1b) comme le coup de chaleur. (f:2a) Pour prévenir ce problème, il faut limiter l'activité physique pendant les heures les plus chaudes de la journée, (f:2b) porter un chapeau ou une casquette, (f:2c) boire beaucoup d'eau et (f:2d) consommer du sel. (f:3) Le désir d'obtenir un bronzage rapide peut jouer de mauvais tours. (f:4a) Mieux vaut s'exposer graduellement au soleil (f:4b) après s'être enduit d'une crème solaire adéquate et (f:4c) en portant des lunettes solaires de bonne qualité.

Nous avons vu que notre système de génération suppose une prise en compte de tous les aspects fonctionnels de la production linguistique. L'input au processus de génération doit donc obligatoirement contenir des éléments tant idéationnels et interpersonnels que textuels[2]. Si l'on observe nos deux textes de manière multifonctionnelle, c'est-à-dire, selon les trois métafonctions idéationnelle, interpersonnelle et textuelle, il apparaît que le contenu idéationnel et interpersonnel est pratiquement équivalent pour les deux langues (l'input sera donc pour ainsi dire équivalent). Cependant, la répartition des phrases, la progression thématique, le choix des conjonctions, etc., sont différents dans les deux textes ; ces phénomènes sont tous d'ordre textuel. Ces divergences devraient être contrôlées par le planificateur de texte pendant le processus de génération. Ceci soulève deux questions. Premièrement, les propositions faites en matière de planification de texte sont-elles suffisantes pour le genre de contrôle que nous demandons ici ? Deuxièmement, est-il possible de construire des bases textuelles multilingues ?

Contraintes textuelles sur la lexicogrammaire

Le point de départ de nos investigations textuelles a été une architecture telle que celle suggérée dans Bateman *et al.* (1991a) et Hovy *et al.* (1992). Nous avons également pris en compte le type et le genre de texte, ce qui permet de se concentrer sur les propriétés particulières de nos textes pris comme exemples et nous avons voulu faire une première application des réseaux multilingues au niveau discursif (voir aussi, Bateman *et al.* (à paraître)).

Le problème général de l'application des principes décrits dans les premiers modèles discursifs, s'est posé en termes de distance entre le niveau textuel très général et le degré de détail des divergences textuelles dans les textes mêmes. Il ne suffit pas, en effet, de donner une description des fonctions communicatives et des relations rhétoriques qui sont incluses dans un texte pour en saisir la structure discursive complète. Une brève analyse des textes en ces termes l'illustre clairement.

Dans un premier temps, comparons les fonctions communicatives des deux textes, telles qu'elles sont données par le **potentiel structurel générique**. Cette notion a été introduite par Hasan (1978) sous le nom de *Generic Structure Potential* (GSP). Il

2. L'input au processus de génération dans le système PENMAN se fait sous forme de SPL (*Sentence Planning Language*) développé par Kasper (1989). Ce mode de représentation a été appliqué à la traduction automatique (Bateman *et al.* 1989 ; Bateman 1992) et à la génération multilingue (Bateman *et al.* 1993).

s'agit de la description du texte en termes d'étapes en fonction desquelles le discours est organisé. Elles sont très similaires aux schémas introduits par McKeown (1985). Il apparaît que les GSP des deux textes sont les mêmes, ce qui est prévisible pour des textes de même type et qui ont le même registre. Il s'agit en effet de textes présentant des conseils généraux et il est typique pour ce genre de texte de présenter les problèmes avec une série de solutions. La figure 3 est une illustration du GSP commun, qui peut être considéré comme un premier candidat pour une définition multilingue de ce qu'est un *même* texte.

Étapes dans le GSP	
1	Problème
2	Solution(s)
3	Problème
4	Solution(s)

FIGURE 3 : Potentiel structurel générique.

Une seconde façon d'aborder l'organisation générale d'un texte est d'en donner la structure rhétorique, telle que définie par Mann et Thompson (1987) dans leur théorie de structure rhétorique *(Rhetorical Structure Theory – RST)* et développée dans Hovy *et al.* (1992). Il s'agit d'une théorie descriptive visant à établir la structure d'un texte à l'aide des relations qui existent entre ses parties ; relations qui sont définies en termes fonctionnels. Il est intéressant de noter qu'il existe en général une corrélation étroite entre le GSP d'un texte et son organisation rhétorique. La structure générale de notre texte exemple relie en effet les différents problèmes présentés par le lien JOINT (ici deux problèmes liés à l'exposition au soleil) ; la solution à ces problèmes est représentée par le lien SOLUTIONHOOD. Les seules différences au niveau des relations rhétoriques apparaissent dans les dernières phrases du texte. Dans le texte néerlandais, n:6, n:7a et n:7b sont liés par la relation JOINT, alors qu'en français, les propositions correspondantes[3] f:4a et f:4b sont reliées par SEQUENCE et f:4ab et f:4c par CIRCUMSTANCE. Cette seule différence sera exprimée au niveau de la grammaire par le choix de conjonctions différentes ; pour le reste les textes ont la même organisation rhétorique. Ceci est illustré dans la figure 4.

3. Il faut noter que la RST d'un texte se base sur les propositions et non sur les phrases. Nous reviendrons sur le découpage différent en néerlandais et en français dans la prochaine section.

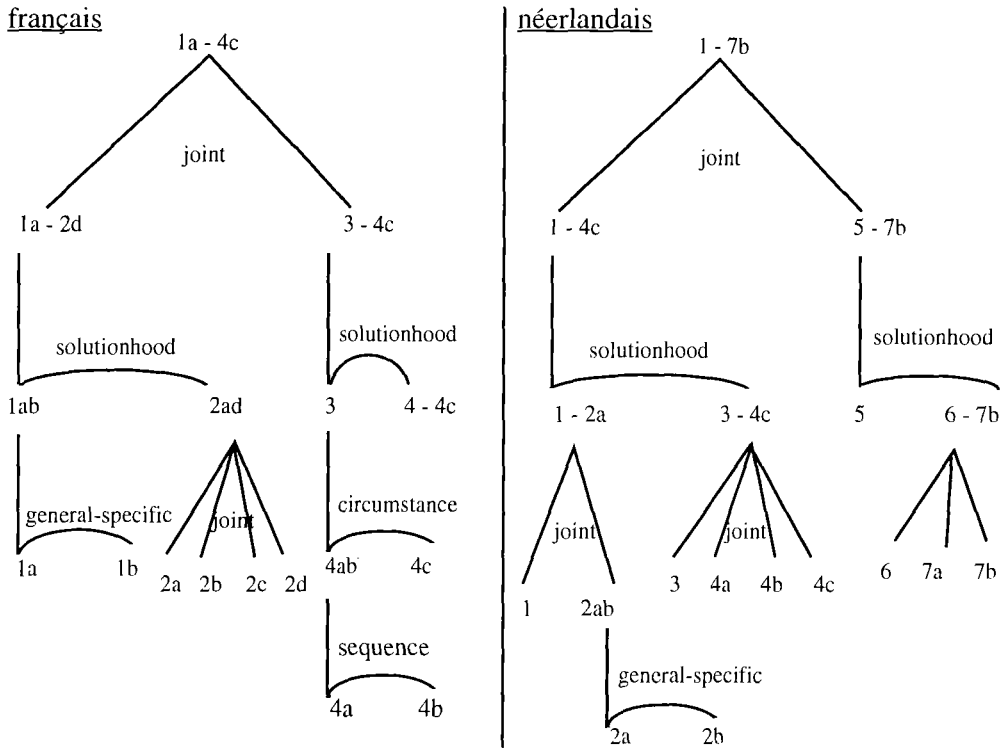


FIGURE 4 : Structures rhétoriques.

Bien que rassurantes pour la description multilingue convergente, ces similarités ne fournissent pas d'explications pour les divergences. Il existe toujours un vide entre l'organisation textuelle globale et la grammaire, ce qui a pour résultat une sous-spécification au niveau textuel plus local qui doit être en mesure de contrôler les choix textuels grammaticaux. Si nous voulons un tel contrôle, il nous faut les moyens de combler ce vide.

Afin d'améliorer cette situation, nous proposons d'adopter de manière systématique l'architecture proposée par Martin (1992), qui suggère une strate détaillée pour la sémantique discursive reliant ainsi les phénomènes textuels globaux aux questions textuelles locales. En introduisant ce niveau de description, nous répondons à un manque de notre ancienne architecture. En effet, afin de contrôler les choix grammaticaux qui sont motivés textuellement, il faut rendre compte des fonctions qu'accomplissent ces éléments dans le discours ; en d'autres mots, il faut réinterpréter la grammaire comme un *moyen de faire du discours*. Un tel traitement de la fonction doit se situer à une strate intermédiaire entre la grammaire et les niveaux plus abstraits d'organisation textuelle, tels que les relations rhétoriques et les schémas. Ce niveau intermédiaire est nécessaire parce qu'il n'y a pas de corrélation directe entre la structure communicative abstraite et les choix grammaticaux qui décident, par exemple, de la thématisation, de la pronominalisation, de la voix passive ou active, etc., phénomènes qui doivent être interprétés en termes de discours.

Un tel niveau est proposé dans la sémantique discursive de Martin (1992), qui contient un nombre de régions fonctionnelles comparables à celles qui existent dans la grammaire. Ces régions sont au nombre de cinq : NÉGOCIATION, IDENTIFICATION, THÉMATICITÉ, CONJONCTION et IDÉATION. Les principales retombées grammaticales de ces systèmes discursifs sont les suivantes. *Négociation* est de nature interpersonnelle, au niveau de la grammaire, cette région sera réalisée, par exemple, par le système de MODE au travers des traits déclaratif, interrogatif et impératif. *Identification* est un système qui traite de la référence à des participants dans la situation décrite. Les choix grammaticaux contrôlés par cette région se trouvent dans la métafonction textuelle ; il s'agit, par exemple, de la pronominalisation, de la référence catégorielle *vs.* spécifique, etc. Une autre région qui affecte la métafonction textuelle est la *thématicité*. Elle met en relation le déroulement global du texte (*method of development* ou progression thématique) et la sélection individuelle des thèmes dans les énoncés. Le système discursif de *conjonction* traite les relations logiques ; il s'agit des relations entre énoncés ou à l'intérieur d'énoncés complexes, par exemple, les relations temporelles ou causales. Dans la grammaire, ces relations sont exprimées par les relations paratactiques (coordonnées) ou hypotactiques (subordonnées) dans les énoncés ou groupes complexes et par le choix de conjonctions,... Finalement, le système d'*idéation* établit le domaine de l'information idéationnelle qui est à fournir dans le texte ; il est à l'origine de la cohésion lexicale.

Nous avons introduit ces systèmes discursifs dans l'ensemble du planificateur textuel, ceci en vue d'obtenir un degré de contrôle plus étroit de l'organisation textuelle au niveau local et également afin de donner une représentation multilingue de ces ressources, comme c'est déjà le cas pour les composantes grammaticales. Comme ces systèmes sémantiques sont également reliés à l'organisation textuelle « supérieure », le contact entre la grammaire et la conception générale du texte est garanti.

Vers une textualité multilingue

Nous avons vu qu'en commençant par un but communicatif général pour créer des textes appartenant à un certain type générique (dans notre cas, des textes informatifs et de conseils), la planification textuelle peut procéder de manière classique jusqu'à l'élaboration d'une organisation textuelle à la RST. La différence qui apparaît dans la structure rhétorique des deux textes au niveau des trois derniers énoncés se reflète assez directement dans la grammaire par le choix des conjonctions. Pour le reste de la structure globale, il n'y a pas de différences significatives. À l'aide des systèmes discursifs locaux introduits dans la section précédente, nous nous concentrons maintenant sur la réalisation grammaticale de cette structure globale.

Le système de *thématicité* est, nous l'avons vu, responsable de la progression thématique dans un texte. Au niveau de la grammaire, celle-ci sera principalement réalisée par le choix des constituants qui rempliront la fonction de thème (et de rhème) dans chacune des phrases et par l'ordre de ces constituants. On distingue en général trois types de progression thématique (Combettes 1988) :

1. *la progression linéaire :* chaque rhème, dans chaque phrase est « l'origine » du thème de la phrase suivante ;

2. *la progression à thème constant :* le même thème apparaît dans des phrases successives, alors que les rhèmes sont différents ;
3. *la progression à thèmes dérivés :* les thèmes sont issus, dérivés, d'un « hyperthème ».

D'habitude, les trois types de progressions se combinent à l'intérieur d'un même texte ; c'est également le cas pour nos deux textes. Toutefois comme l'illustrent les figures 5 et 6, les progressions thématiques ne sont pas les mêmes pour les deux langues.

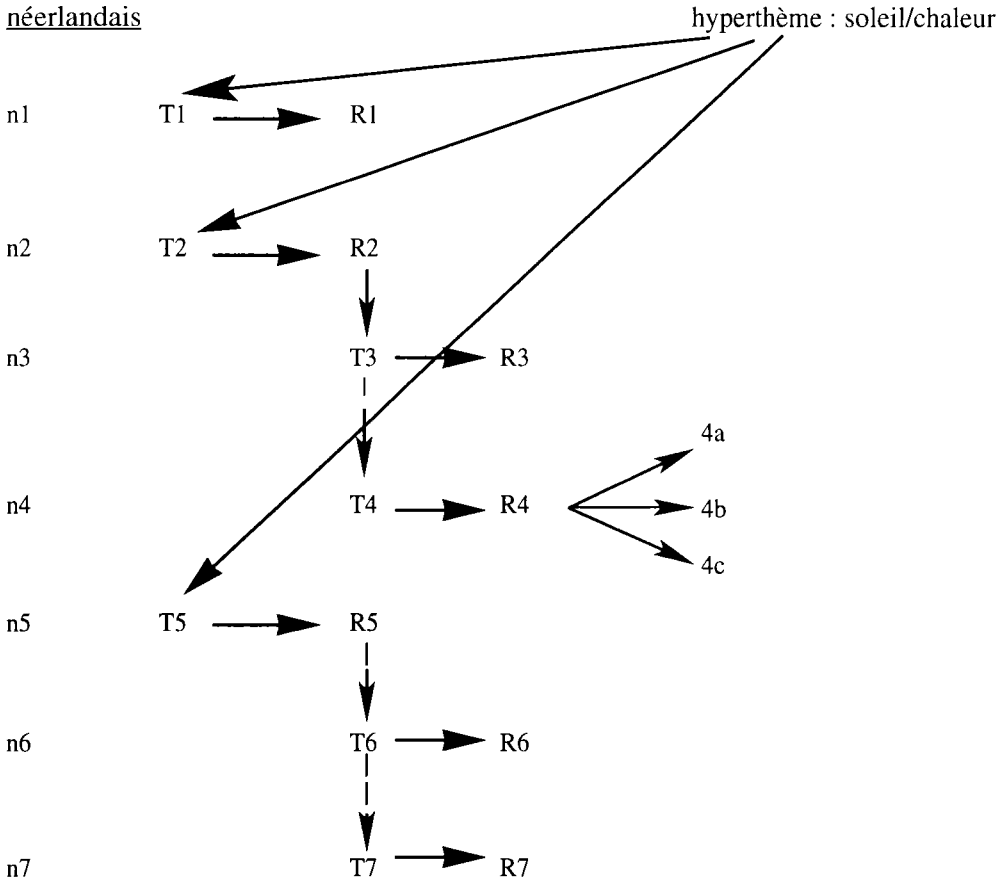


FIGURE 5 : Structure thématique du texte en néerlandais.

Dans le texte néerlandais, les thèmes des phrases n1 et n2 sont dérivés de l'hyperthème, le passage de n2 à n3 se fait ensuite par progression linéaire, et celui de n3 à n4 par progression à thème constant (nous entendons également par thème constant, les thèmes qui expriment une constance au niveau idéationnel ou interpersonnel ; ceux-ci sont indiqués par une flèche interrompue). Le thème de la phrase n5 est à nouveau dérivé de l'hyperthème, après quoi la progression se poursuit de manière linéaire jusqu'à la phrase n6 et par thème constant jusqu'à n7.

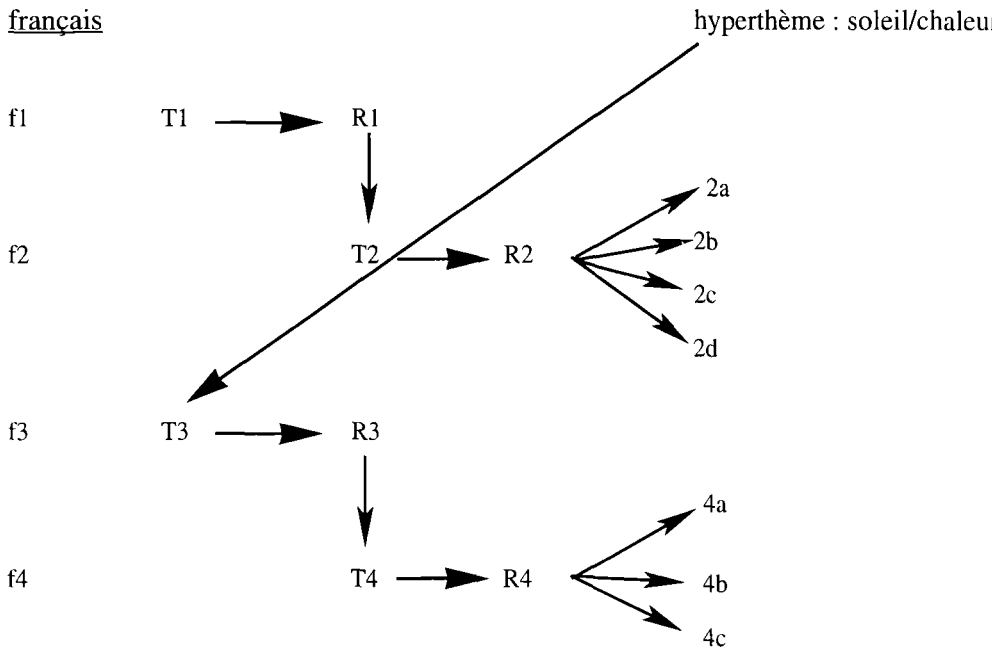


FIGURE 6 : Structure thématique du texte français.

Le texte français est introduit par un nouveau thème et se poursuit par une progression linéaire, l'hyperthème est repris dans la phrase f3, et ensuite il y a à nouveau progression linéaire jusqu'à la phrase f4.

Ce qui frappe en premier lieu, lorsque l'on compare le développement thématique des deux textes, c'est la divergence dans le nombre d'étapes, répercutée dans le découpage phrastique des textes : le texte néerlandais utilise sept phrases, là où le texte français n'en a que quatre pour exprimer le même contenu. Nous reviendrons sur cette différence qui est principalement sous le contrôle du système discursif de *conjonction*. Outre le nombre différent de thèmes, les éléments remplissant la fonction de *thème* sont également différents, ce qui se répercute dans les progressions thématiques différentes. En effet, le texte néerlandais introduit l'hyperthème (introduit par le titre) à trois reprises, et surtout les thèmes des deux premières phrases sont dérivés de l'hyperthème. Ceci est typique des textes qui visent d'abord à être explicites (où l'on rappelle constamment et explicitement de quoi il s'agit), comme c'est le cas dans des textes didactiques comme celui-ci. Les progressions linéaires viennent à la suite des thèmes dérivés. Remarquons aussi que les progressions à thème constant ont lieu là où le français a préféré les constructions à rhèmes multiples (un seul thème pour plusieurs rhèmes dans une phrase complexe). Le texte français joue en effet beaucoup moins sur la répétition. De prime abord, le texte est introduit par un nouveau thème – pas de reprise de l'hyperthème – c'est-à-dire que l'information que *les brûlures sont un risque lié au soleil et à la chaleur* est présupposée activée chez le lecteur dès qu'il lit le titre du passage, alors que dans le texte néerlandais, cette même

information est présentée comme rhématique. La progression se poursuit ensuite de manière linéaire par une construction à rhème multiple ; l'hyperthème est repris comme *thème* de la troisième phrase (pour « relancer » le texte) après quoi il y a à nouveau progression linéaire avec rhèmes multiples. Tout en étant moins explicite, ce développement thématique semble plus directement lié aux buts communicatifs généraux exprimés dans le GSP du texte.

Mais revenons-en maintenant au découpage phrastique différent des deux textes. Nous avons vu que la réalisation grammaticale des énoncés en phrases simples ou en phrases complexes (et donc aussi le choix des conjonctions) était sous le contrôle du système discursif de *conjonction*. Dès qu'il y a plusieurs procès en jeu, il faut faire le choix de les exprimer en phrases simples ou en phrases complexes. Dans le premier cas, on parlera de lien *cohésif* entre les procès, dans le deuxième cas de lien *paratactique* ou *hypotactique* selon le type de dépendance exprimée entre les deux énoncés. Ces choix seront contraints par le type de relations existant entre les énoncés apparaissant dans la structure rhétorique des textes. L'aboutissement du découpage en phrases est exprimé dans la structure thématique des textes. Ainsi les systèmes discursifs de *conjonction* et de *thématicité* semblent véritablement jouer un rôle de trait d'union entre la structure rhétorique (encore assez abstraite) d'un texte et sa structure thématique (proche de la réalisation grammaticale) ; mais la façon dont ces « traits d'union » seront mis en œuvre sera spécifique à chaque langue. Pour preuve, les structures rhétoriques du texte néerlandais et français sont pratiquement similaires, alors que les structures thématiques sont fort divergentes. Lorsque l'on compare ces deux types de structures pour chacun des textes, il apparaît qu'il y a une assez grande convergence pour le texte néerlandais : chaque nœud de la structure rhétorique semble correspondre à une étape dans la structure thématique[4]. Ceci est à nouveau typique pour le genre didactique de ce texte, où chaque énoncé tend à être réalisé par une phrase séparée. En français, ceci n'est pas le cas, il y a seulement convergence au niveau du sommet de la structure rhétorique, qui exprime le découpage général du texte et est directement lié au GSP.

Un dernier point de divergence que l'on peut brièvement aborder ici concerne la réalisation des formes impersonnelles qui sont régies par le système discursif de *négociation*. Au niveau sémantique, ce système fait la distinction entre les différents actes de paroles qui seront réalisés grammaticalement par le choix des modes impératif, déclaratif, interrogatif, etc. Les actes de paroles présents dans nos exemples sont de deux types : donner des informations et donner des ordres (sous forme de conseils). Pourtant, il n'y a pas de forme impérative (la réalisation typique des ordres) dans aucun des deux textes, seulement des formes déclaratives. Ceci est une conséquence du registre du texte qui est impersonnel (voir aussi Butler 1988). Les ordres sont donc plutôt exprimés sous forme de conseils par des formes linguistiques impersonnelles. Ces constructions sont spécifiques à chaque langue. En néerlandais, on utilisera plus volontiers la forme passive (n3), le pronom impersonnel *men* (qui est beaucoup plus formel en néerlandais qu'en français) avec un verbe modal (n7), et des constructions à sujets syntaxiques vides (n4 et n6). En français, seul ce dernier type de construction est utilisé (f2 et f4).

4. D'autres contraintes grammaticales, voire stylistiques, peuvent également intervenir. Ainsi, le lien JOINT ne sera pas exprimé par un lien cohésif si le sujet grammatical et le type de procès sont les mêmes dans les deux énoncés. Dans ce cas, le lien sera plutôt exprimé par une ellipse et une conjonction paratactique.

Conclusion

Dans cet article, nous avons exploré les possibilités d'un système de génération multilingue à typologies fonctionnelles. Nous avons vu qu'un tel système peut rendre compte des convergences qui existent entre les langues et également des divergences, et ceci aux différentes strates du système linguistique. Notre attention s'est concentrée sur les aspects textuels qui interviennent dans la génération de textes, aspects qui doivent être pleinement intégrés dans le système, si nous voulons générer de véritables textes et non seulement de simples suites de phrases. Cette textualité est bien souvent divergente dans les différentes langues, et surtout elle sera exprimée par des structures lexicogrammaticales différentes. Le passage de la strate sémantique (discursive) à la strate lexicogrammaticale nécessite donc l'instauration d'un niveau sémantique intermédiaire afin de contrôler les choix textuels à ce niveau local. Un tel rôle peut être joué par les systèmes discursifs tels que ceux introduits par Martin (1992).

Le travail présenté dans cet article est, bien sûr, encore à son stade de développement. Les hypothèses émises doivent être confrontées à un nombre de textes beaucoup plus grand, de genre et de registre différents.

Remerciements

Ce travail repose largement sur le travail accompli dans le projet KOMET à GMD/ IPSI, Darmstadt. Je remercie mes collègues John Bateman et Elke Teich pour les nombreuses discussions et remises en question. Je remercie également Alain Polguère pour la relecture du manuscrit et ses nombreuses suggestions.

Cette recherche est partiellement financée par le projet de recherche fondamentale ESPRIT - Dandelion, n° 6665.

Références

BATEMAN, J. A. (1992) : « Towards Meaning-based Machine Translation », *Machine Translation*, vol. 6, n° 1, pp. 1-37.

BATEMAN, J. A., KASPER, R. T., MOORE, J. D. et R. A. WHITNEY (1989) : « A New View on the Translation Process », *Proceedings of the European Chapter of the Association for Computational Linguistics*, Manchester, pp. 282-290.

BATEMAN, J. A., MAIER, E. A., TEICH, E., et L. WANNER (1991) : « Towards an Architecture for Situated Text Generation », *Proceedings of the International Conference on Current Issues in Computational Linguistics,* Penang, Malaysia, pp. 336-350.

BATEMAN, J. A., MATTHIESSEN, C. M. I. M., NANRI, K. et L. ZENG (1991) : « The Re-use of Linguistic Resources Across Languages in Multilingual Generation Components », *Proceedings of the 1991 International Joint Conference on Artificial Intelligence,* Sydney, Australia, Morgan Kaufmann Publishers.

BATEMAN, J. A., DEGAND, L. et E. TEICH (1993) : « Multilingual Textuality: some Experiences from Multilingual Text Generation », *Proceedings of the Fourth European Workshop on Natural Language Generation,* Pisa, Italy, April 1993.

BATEMAN, J. A., MATTHIESSEN, C. M. I. M. et L. ZENG (à paraître) : *A General Architecture for Multilinguality in Natural Language Processing.*

BUTLER, C. S. (1988) : « Politeness and the Semantics of Modalised Directives in English », Benson, J. D., Cummings, M. J., Greaves, W. S. (dir), *Linguistics in a Systemic Perspective,* Amsterdam, Benjamins, pp. 119-154, Current Issues in Linguistics Theory.

COMBETTES, B. (1988) : *Pour une grammaire textuelle. La progression thématique,* De Boeck-Wesmael, Bruxelles, Duculot, Paris, Gembloux.

DEGAND, L. (1993) : *Towards a Systemic Functional Grammar of Dutch for Multilingual Text Generation,* Technical Report, Komet Project, GMD/IPSI, Darmstadt, [Forme abrégée dans *Proceedings of the Fourth European Workshop on Natural Language Generation,* Pisa, Italy, April 1993].

DEGAND, L. (à paraître) : *Dutch Grammar Documentation,* Technical Report, Komet Project, GMD/IPSI, Darmstadt.

HALLIDAY, M. A. K. (1978) : *Language as Social Semiotic,* London, Edward Arnold.

HALLIDAY, M. A. K. (1985) : *An Introduction to Functional Grammar,* London, Edward Arnold.

HASAN, R. (1978) : « Text in the Systemic-Functional Model », Dressler, W. (dir), *Current Trends in Text Linguistics,* Berlin, de Gruyter, pp. 228-246.

HOVY, E., LAVID, J., MAIER, E., MITTAL, V. et C. PARIS (1992) : « Employing Knowledge Resources in a New Text Planner Architecture », Dale, R., Hovy, E., Rösner, D., Stock, O. (dir), *Aspects of Automated Natural Language Generation, Proceedings of the 6th International Workshop on Natural Language Processing,* Berlin, Springer, pp. 57-72.

KASPER, R. T. (1989) : « A Flexible Interface for Linking Applications to PENMAN's Sentence Generator », *Proceedings of the DARPA Workshop on Speech and Natural Language.*

MANN, W. C. et S. A. THOMPSON (1987) : *Rhetorical Structure Theory: A Theory of Text Organization,* USC Information Sciences Institute Technical Report RS-87-190.

MARTIN, J. R. (1992) : *English Text: System and Structure,* Amsterdam, Benjamins.

MATTHIESSEN, C. M. I. M. (1992) : *Lexicogrammatical Cartography: English Systems,* University of Sydney, Australia.

MATTHIESSEN, C. M. I. M. et J. A. BATEMAN (1991) : *Text Generation and Systemic-functional Linguistics: Experiences from English and Japanese,* London, Pinter Publishers.

MCKEOWN, K. R. (1985) : *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Texts,* Cambridge, Cambridge University Press.

5

Acquisition et préparation des ressources textuelles pour le TAL

Susan WARWICK-ARMSTRONG

ISSCO, Université de Genève, Genève, Suisse

Arrière-plan

Pour un nombre croissant de chercheurs, l'accès électronique à des collections étendues de textes et à leurs traductions est devenu une ressource essentielle pour l'analyse du langage et les études sur la traduction. Des méthodes empiriques et statistiques sont en train de se développer pour organiser les données et élaborer ainsi des modèles plus adéquats de la structure et de l'usage des langues naturelles. Elles permettent déjà le marquage des textes en parties du discours, la prédiction de séquences de mots, la reconnaissance des collocations et l'alignement automatique des phrases et de leur traduction. Elles offrent ainsi un point de départ pour des études plus approfondies et pour des applications pratiques dans des domaines variés tels que la lexicographie, la reconnaissance vocale et la TA.

Cet intérêt récent et croissant pour les études basées sur les corpus est en quelque sorte une réminiscence des méthodes empiriques et statistiques en vogue dans les années 50, puisque les premiers travaux en TA – une des premières applications de la linguistique informatique – étaient liés aux problèmes de décryptage (Weaver 1949). À cette époque, cependant, les ressources informatiques étaient loin d'être adéquates et les ressources textuelles, indispensables aux modèles statistiques, inexistantes. Il est clair aujourd'hui que les avances technologiques en matière de puissance informatique, ainsi que le nombre croissant de textes disponibles, ont certainement favorisé le retour de cette approche.

Un autre facteur important a contribué à l'intérêt des méthodes orientées données :

les systèmes basés sur les règles n'ont pas produit les résultats attendus. Ils ne sont pas non plus directement source d'espoir pour l'avenir.

Il va sans dire que cette nouvelle orientation a aussi suscité un certain nombre de critiques – un débat qui peut être caractérisé par les termes « basé sur les statistiques *vs* basé sur les règles », « empirique *vs* rationaliste ». Néanmoins, la plupart des chercheurs engagés dans les études basées sur le corpus ne considèrent pas les statistiques comme la solution unique de description du langage, mais plutôt comme une méthode qui fournit des solutions partielles aux différents problèmes du TAL. Une des directions importantes dans ce domaine est d'ailleurs l'intégration de modèles de probabilité et de systèmes basés sur les règles et, inversement, l'enrichissement des modèles statistiques du langage par des informations linguistiques plus traditionnelles.

Quoi qu'il en soit, les travaux récents montrent que les nouvelles méthodes orientées vers les données offrent des solutions potentielles aux problèmes clés de la linguistique informatique :

- acquisition — identification et codage de toutes les informations nécessaires ;
- couverture — rendre compte de tous les phénomènes d'un domaine spécifique, d'une collection de textes ou d'une application ;
- robustesse — concilier les données réelles qui peuvent être non grammaticales ou simplement non prises en compte par le modèle ;
- extensibilité — application du modèle et des données à un nouveau domaine, un nouvel ensemble de textes, une nouvelle problématique, etc.

Quoique ces problèmes ne soient pas neufs, l'accès à des ressources étendues de textes permet d'explorer de nouvelles directions de recherche, sources potentielles de progrès tant en recherche fondamentale que pour l'aide à la résolution de problèmes très pratiques tels que la construction de dictionnaires, la classification des noms propres et des mots inconnus ainsi que l'identification des phrases nominales et autres collocations.

Dans la suite, nous présenterons quelques questions générales concernant l'acquisition et la préparation des corpus et nous reprendrons un certain nombre d'activités qui concernent les collectes de données en Europe et en Amérique du Nord.

Disponibilité des données textuelles

Dans la dernière moitié des années 80, quand l'intérêt pour les méthodes statistiques et les travaux basés sur les corpus émergeait au sein de la communauté de la linguistique informatique, il n'existait que très peu de données disponibles pour servir les buts de la recherche. Cette situation contraste avec celle que connut la communauté de la parole où les modèles de probabilité et les méthodes statistiques sont devenus le standard et où le rassemblement des données a donc été considéré comme une partie intégrante de n'importe quel projet (voir Church et Mercer (1993) pour une discussion sur le développement des méthodes statistiques dans le domaine de la parole et ses effets sur les travaux en linguistique informatique et Liberman (1992), pour quelques exemples de l'utilité de corpus pour la communauté de la parole).

Les plus vieux corpus pour l'anglais, tels que le *Brown corpus* (Francis et Encera 1982), étaient relativement petits et d'autres comme le *Birmingham corpus* (Sinclair 1987) n'étaient pas accessibles publiquement. En Europe continentale, où ce nouvel intérêt pour les études basées sur les corpus vient seulement d'émerger récemment, la situation est similaire : les textes contenus dans les centres linguistiques nationaux sont soit trop importants pour le chercheur individuel, soit peu utiles pour les méthodes actuelles, soit simplement non disponibles pour le public.

Quoiqu'il y ait un potentiel important de données sous forme électronique, la communauté scientifique ne dispose que d'une petite portion de ce matériel. Les textes sont disséminés dans divers centres privés et les détenteurs de données sont souvent des maisons d'édition qui n'ont pas les droits de détention et de distribution. Cette distinction entre détenteurs et possesseurs est également apparente dans les grandes organisations où les services techniques qui gèrent les archives sont bien séparés des autres départements. La simple localisation des données sous forme électronique est souvent un problème en lui-même une fois le texte publié.

Le manque de matériaux textuels appropriés (en taille et en type de données) a limité les travaux de recherche de différentes manières. Tout d'abord, il y a très peu de matériel disponible pour les langues autres que l'anglais : la recherche s'est donc concentrée sur la langue anglaise et les méthodes ont été configurées pour tirer parti de certains phénomènes spécifiques à cette langue, comme l'ordre fixe des mots et la morphologie limitée. Il convient de voir comment elles peuvent être étendues aux autres langues. Ensuite, en raison de la disponibilité publique d'un seul corpus[1], les études de traduction se sont concentrées sur une paire de langues et sur un type de textes. Enfin, la nature privée de la plupart des données en usage a conduit à la duplication du travail plutôt qu'à son partage et à l'extension des résultats.

Heureusement, cette situation est en train de changer et c'est ce progrès que nous voudrions illustrer ici. Nombre d'initiatives (voir infra) ont contribué à mettre en évidence le désir et le besoin d'un accès public aux données et à démontrer l'intérêt d'une coopération en matière d'acquisition et de préparation de ces ressources. Les données qui furent disponibles au travers d'initiatives et de quelques efforts individuels ont permis une collection plus ou moins *ad hoc* de données. La Communauté a travaillé dans l'idée que quelques données étaient préférables à pas de données du tout et que le plus grand nombre de données était le mieux. Néanmoins, dès qu'un grand nombre de textes seront accessibles, se posera la question de la manière de construire un corpus *équilibré* et *représentatif* – ce que Walker (1991) a intitulé l'*écologie du langage*. Dans le but de fournir une couverture adéquate du langage à un moment donné et pour un domaine donné, il faut considérer des critères tels que le style, le registre, le type de texte, la fréquence, etc. (Biber 1993).

Une question importante que toute entreprise de collecte de données doit se poser est comment protéger les intérêts des auteurs des textes – une question d'un intérêt critique dans cette nouvelle ère électronique. Alors que les textes sont normalement acquis et consultés pour leur valeur informative et dérivative, leur utilisation dans les études basées sur le corpus est bien différente. L'intérêt réside dans l'usage

1. Les débats parlementaires canadiens, connus sous le nom de *corpus Hansard*.

de la langue (c'est-à-dire les expressions telles qu'elles apparaissent dans les textes) plutôt que dans le contenu d'un document donné. Cette idée des textes, mesurés en kilo-octets (Ko) plutôt qu'en contenu, est souvent difficile à expliquer au détenteur des données qui considère les textes en termes de droits d'auteur, de valeur artistique ou dérivative, ou, plus simplement, comme une source potentielle de revenus. Et contrairement au passé où l'environnement de la recherche consistait seulement à accéder à une bibliothèque richement dotée (ou à un système de prêts interbibliothèques) et à des ressources informatiques adéquates, pour des études basées sur le corpus, chaque groupe de recherche doit avoir une copie personnelle de tout le matériel.

À la lumière de cette situation, les entreprises de collection de données établissent des accords formels tant avec les fournisseurs de données qu'avec les utilisateurs potentiels de ces données. Dans le cas des organisations décrites ci-dessous, chaque demandeur de données doit signer un accord de non-diffusion des données et de respect de toutes les restrictions stipulées par les fournisseurs. Quoique beaucoup de questions d'accès, de droits d'auteurs et de protection de données en général (comme le matériel privé, les collections qui indiquent des sources potentielles de revenus, etc.) doivent toujours être clarifiées, ces accords fournissent une base légale pour se protéger d'abus d'utilisation.

Initiatives de collection de données

Nous décrirons maintenant brièvement quelques activités récentes concernant la collection de textes et leur diffusion. Nous commencerons par des efforts privés et ensuite nous regarderons leur suivi dans les projets officiels qui assureront une base de structure plus solide.

ACL/Data Collection Initiative

La première de ces initiatives, l'*ACL/Data Collection Initiative (ACL/DCI)* a été fondée en 1989 par l'*Association for Computational Linguistics.* Sous ses auspices s'est développée une association scientifique sans but lucratif pour la supervision, l'acquisition et la préparation d'un large corpus de textes à mettre, sans paiement de redevances, à la disposition de chercheurs scientifiques. Le travail d'acquisition fut réalisé par des volontaires, celui de nettoyage et de préparation (sous forme SGML) du matériel par une petite équipe (Liberman 1989).

En 1991, l'ACL/DCI produisit et diffusa son premier CD-ROM et une centaine de sites travaillent actuellement avec ces données. Le disque contient plus de 600 Ko de données, la plupart anglaises et américaines et inclut, entre autres, une collection étendue d'articles de journaux du *Wall Street Journal*, un dictionnaire anglais offert par les éditions Collins et quelques données annotées grammaticalement du projet *Penn Tree Bank* (Marcus *et al.* 1993) ; un second CD-ROM est actuellement en préparation.

Initiative européenne de corpus

Une initiative similaire a été établie en 1991 : l'*European Corpus Initiative* (Thomp-

son 1992) pour l'acquisition d'un large corpus multilingue destiné aux travaux de recherche en Europe. L'accent a été mis principalement sur le rassemblement des textes en langue non anglaise afin de fournir une base aux chercheurs de tous les pays européens désireux de travailler sur leur propre langue nationale. Un but additionnel a été d'acquérir un ensemble de textes en parallèle (les textes et leurs traductions) vu l'importance de la production de documents multilingues en Europe et l'intérêt porté aux études sur la traductions.

Un grand nombre de données ont maintenant été recueillies dans la plupart des langues européennes, avec au moins 5 millions de mots de textes pour chacune des langues majeures. Une variété de corpus parallèles a également été acquise auprès des organisations internationales et des banques suisses (respectivement, en anglais, en français, en espagnol et en anglais, en français, en allemand). Les textes sont disponibles sur CD-ROM[2].

Établissement de répertoires de textes

Les deux initiatives mentionnées ci-dessus ont été mises en exergue : elles ouvrent une voie nouvelle afin de subvenir aux besoins des chercheurs de la communauté linguistique informatique. Ces efforts bénévoles sont maintenant lentement suivis par des projets officiels, qui devraient établir une base de financement et une infrastructure propre pour développer ces ressources à long terme.

« Linguistic Data Consortium »

Aux États-Unis, le gouvernement américain a fondé le *Linguistic Data Consortium* (LDC), reconnaissant aussi la nécessité de remplacer les efforts informels, pour une large part, par une base structurelle solide. Un autre but était de rendre disponibles les ressources à tous les chercheurs, et pas seulement ceux qui travaillent dans de grands laboratoires privés (qui ont déjà accès aux données internes et à un budget pour acquérir et préparer des collections privées). Le LDC fut fondé en 1992 avec un fonds initial octroyé par l'*Advanced Research Projects Agency* (ARPA) pour acquérir, préparer et diffuser le matériel pour la communauté scientifique (Liberman 1992). En moins d'un an, le LDC a produit près de 100 CD-ROMs. Il travaille actuellement à l'acquisition de données supplémentaires. Un des buts principaux pour l'année prochaine sera l'acquisition de textes multilingues qui aident la traduction automatique et favorisent d'autres activités[3].

Efforts multinationaux

En Europe, où l'environnement multilingue pose des problèmes spécifiques à une action centralisée dans ce domaine, le travail a débuté par la définition d'un cadre pour des actions supplémentaires en vue de constituer des ressources textuelles en Europe.

---

2. Le CD-ROM a été réalisé par le LDC ; la distribution est assurée en Europe par ELSNET (email : elsnet@cogsci.edinburgh.ac.uk) et aux États-Unis par LDC.
3. Contacter : The Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305 ; email : ldc@unagi.cis.upenn.edu.

Une étude initiale de faisabilité a été réalisée dans le cadre du projet *Network for European Corpora* (NERC). Un projet de suivi dont l'intention est d'établir l'infrastructure appropriée pour les collections de textes en Europe et la diffusion de données, a démarré au début de cette année dans le cadre du projet fondé par le CEC, RELATOR[4].

Un autre projet, a commencé cette année (suivi par ECI) : il consiste dans la collecte et la préparation d'un large corpus multilingue (Thompson 1993). Le corpus sera constitué d'un ensemble de documents comparables plurilingues, dans au moins six langues européennes (articles de journaux dans le domaine de la finance) et d'un corpus parallèle multilingue dans l'ensemble des neuf langues (essentiellement issu des *Publications Officielles de la Communauté européenne*)[5]. Ce projet permettra donc d'offrir des textes similaires dans toutes les langues.

Préparation de données

Si les problèmes de base relatifs à l'acquisition et la négociation des droits de diffusion sont nombreux, rendre les données utilisables requiert souvent un bon nombre d'efforts pour « nettoyer » et reformater les données. Avoir simplement les données sous une forme électronique ne suffit pas nécessairement, quoique dans le futur, avec l'extension de la publication électronique de documents et l'établissement de standards de marquage et de codification, ce problème puisse disparaître. Étant donné le temps consacré actuellement à cette tâche dans chaque activité de collecte de corpus – une situation à envisager pour peut-être une décennie sinon plus – ce n'est pas une tâche qui peut être sous-estimée. Des textes plus anciens qui ne sont préparés que pour l'impression sont généralement stockés sur bandes magnétiques dans un format complexe et non documenté ; la correspondance entre la structure logique du texte et la structure typographique est souvent difficile à établir.

Les collections étendues de textes des Nations Unies, récemment acquises par LDC en sont un bon exemple (Graff 1993). Les documents en anglais, en français et en espagnol sont archivés sur bandes magnétiques issues du système informatique Wang, une solution efficiente pour le stockage, mais non pour l'extraction de tous les fichiers. Extraire les données des textes à partir des bandes magnétiques a demandé un effort considérable (avec l'aide de la société Wang elle-même) pour décrypter le code de caractères spécifique au système, les codes de contrôle et de format et la structure du fichier. La mise en concordance de textes parallèles n'a pu se faire que de manière semi-automatique, étant donné les conventions de noms de fichiers *ad hoc* entachées de nombreuses erreurs humaines lors de l'introduction. En effet, les anciens systèmes n'ont pas prévu une telle application : le langage de marquage a été développé à des fins d'affichage physique, plutôt qu'en vue de représentation logique des informations.

Outre ces considérations concernant les schémas de marquage pour le formatage de textes, il existe un débat sur les standards pour l'annotation par exemple, le marquage

4. NERC et RELATOR sont subventionnés par la CEE, DG-XIII, Luxembourg ; contacter : Roberto Cencioni, Blvd. Jean Monnet, 2920 Luxemboug ou Nono Varile, email : M444@eurokom.ie.
5. Ce projet fait partie du *CEC International Scientific Cooperation Program.*

additionnel ajouté aux données. La relative limitation des données vraiment disponibles et des recherches actuelles sur les informations qui peuvent être identifiées dans les textes, explique que chaque projet de corpus ait adopté des conventions internes, par exemple pour le marquage du mot et de la phrase, des parties du discours et de la structure de la phrase. Aussi longtemps que le marquage est clair, bien documenté, sans ambiguïté, et aisé à convertir pour une implémentation sur la machine locale, différentes conventions peuvent suffire pour ces différentes tâches. Toutefois, comme l'information associée aux données devient plus complexe, les standards ou les conventions adoptées deviennent problématiques. Un des plus importants projets internationaux, le *Text Encoding Initiative* s'attache à définir un ensemble de règles directrices des codes pour tous types de marquage de textes avec une attention spéciale aux besoins complexes des chercheurs.

En travaillant sur des textes dans un environnement multilingue, surviennent également un certain nombre de questions qui n'apparaissent pas nécessairement dans un environnement unilingue. L'interprétation d'un symbole donné peut être différente pour une langue donnée (les conventions du code alphabétique, par exemple). Le problème devient plus complexe quand une information associée aux données textuelles n'a pas d'équivalent dans une autre langue. Des études sont en cours afin de déterminer dans quelle mesure les systèmes de marquage essentiellement basés sur l'anglais peuvent être étendus vers les langues européennes qui affichent une gamme plus large de phénomènes morpho-syntaxiques (Monachini et Ostling 1992).

Comme de plus en plus de données, corrigées manuellement, sont préparées à l'aide d'un marquage linguistique sophistiqué, tâche longue et coûteuse, la standardisation de l'annotation devient une question primordiale. Afin de promouvoir le partage des ressources et la comparaison de résultats, des schémas de codification communs deviennent un but souhaité. Un des projets européens importants MULTEXT, qui a pour but de rendre un large corpus multilingue, (partiellement) validé manuellement, disponible avec des annotations relatives à la structure logique du texte, au marquage des phrases, et à l'alignement des textes parallèles, s'attachera de manière systématique à ce problème.

Ces quelques remarques sur la préparation des textes et leur marquage mettent en exergue un bon nombre de questions, qui deviennent cruciales avec l'augmentation des données disponibles dans une gamme de langues différentes. Les jeux de caractères et les codes de formatage de textes devront être standardisés afin de permettre les échanges internationaux de données. Pour des marquages de plus haut niveau, il est peut-être prématuré de prévoir une standardisation dans ce domaine. Comme les nouvelles méthodes évoluent et sont appliquées aux données et comme les résultats sont partagés, de nouveaux standards et conventions émergeront certainement.

Références

BIBER, D. (1993) : « Using Register-Diversified Corpora for General Language Studies », *Computational Linguistics*, vol. 19 n° 2, pp. 219-242.

CHURCH, K. et R. MERCER (1993) : « Introduction to the Special Issue on Computational Linguistics Using Large Corpora », *Computational Linguistics*, vol. 19 n° 1, pp. 1-24.

FRANCIS, W. et H. KUČERA (1982) : *Analysis of English Usage,* Houghton Mifflin.

GRAFF, D. (1993) : « The UN Multilingual Text Corpus », *LDC Newsletter,* Linguistic Data Consortium, vol. 1 n° 3.

LIBERMAN, M. (1989) : « Text on Tap: The ACL/DCI », *Proceedings of the 1989 DARPA Speech and Natural Language Workshop,* Cape Cod.

LIBERMAN, M. (1992) : *Introduction to the Linguistic Data Consortium,* distribué à COLING, Nantes.

MARCUS, M., SANTORINI, B. et B. MARCINKIEWICZ (1993) : « Building a Large Annotated Corpus of English: The Penn Tree Bank », *Computational Linguistics,* vol. 19 n° 2, pp. 313-331.

MONACHINI, M. et A. OSTLING (1992) : *Morphosyntactic Corpus Annotation - A Comparison of Different Schemes,* Istituto di Linguistica Computazionale, CNR, Pisa, report for NERC project.

SINCLAIR, J. (dir) (1987) : *Looking Up: An Account of the COBUILD Project in Lexical Computing,* Londres, Collins.

THOMPSON, H. (1992) : « European Corpus Initiative », *ELSNEWS,* vol. 1 n° 1, Edinburgh, Center for Cognitive Science.

THOMPSON, H. (1993) : *Multilingual Corpora for Cooperation (MLCC),* Proposal submitted under the LRE program for International Scientific Cognitive Science-operation, Luxembourg, Commission des Communautés européennes.

WALKER, D. (1991) : « The Ecology of Language », *Proceedings of the International Workshop on Electronic Dictionaries,* Japan Electronic Dictionary Research Institute, Tokyo, pp.1-22.

WEAVER, W. (1949) : *Translation,* New York, (memorandum).

6

Une approche par acceptions pour les bases lexicales multilingues

Gilles SÉRASSET et Étienne BLANC

GETA, IMAG, Université Joseph-Fourier et CNRS, Grenoble, France

• *Abstract* •

Many projects are conducted to develop multilingual lexical databases. Some of these projects use an interlingual approach (KBMT-89, EDR, ...), where others choose a bilingual approach (Multilex, ...).

GETA uses an interlingual approach based on acceptions (word-senses) to develop its multilingual lexical database management system : NADIA. With this approach, the interlingual lexicon is the union of the acceptions of the languages that appear in the database.

The interlingual set of relations is freely defined by the linguist, with the exception of the predefined "isa" relation, which is necessary in any case because acceptions of terms from one language don't always have lexical equivalents in another language.

This lexical architecture has no influence on the linguistic content of the monolingual dictionaries, neither on the format adopted by the linguist to code the information.

The NADIA system is designed to be multilingual (the system handles multilingual databases), application independent (databases can be used for several purposes), generic (the linguistic structures are defined by the linguist) and theory independent (many computational formalisms can be used to define the linguistic structures). The system provides many tools (browser, editor, defaulter, coherence checker, ...) to simplify the management of a multilingual lexical database. Moreover, it handles the management of the interlingual dictionary as much as possible.

EDR and KBMT-89 projects (which use a knowledge based approach) are faced with theoretical and methodological problems (concept discrimination, knowledge representation, concept classification, ...). The acception based approach is a good choice to avoid the complexity introduced by knowledge representation problems and to keep advantages of the interlingual approach.

Multilex provides a way for the linguist to define the different linguistic structures, however, the linguist can only use typed feature structures to code his linguistic theory. At GETA we have chosen to provide the linguist with different formalisms (trees, graphs, typed feature structures, sets, ...).

Introduction

Les besoins en ressources lexicales de grande taille pour le traitement automatique des langues naturelles (TALN) en général et la traduction automatique (TA) en particulier, augmentent chaque jour. On considère que ces ressources représentent la partie la plus coûteuse d'un système de TALN. Pour cette raison, on observe un intérêt croissant pour le développement de dictionnaires réutilisables.

Pour développer une base de données lexicale multilingue, deux approches peuvent être utilisées. En premier lieu, *l'approche par transfert,* où les liens entre les langues se font au travers de dictionnaires bilingues et unidirectionnels. Cette approche est utilisée par de nombreux systèmes de TA, ainsi que par certains projets de bases de données, comme notamment, les projets Acquilex et Multilex. En second lieu, *l'approche interlingue,* où les liens entre les langues se font au travers d'un dictionnaire interlingue unique. Une telle approche a été adoptée aux USA par le projet KBMT-89 (*Knowledge Based Machine Translation*), à l'université Carnegie Mellon et par le projet japonais EDR (*Electronic Dictionary Research*).

Le laboratoire GETA (Groupe d'Étude pour la Traduction Automatique) s'intéresse aux problèmes posés par la construction et l'utilisation de bases de données lexicales multilingues, indépendantes d'une théorie linguistique et indépendantes d'une application. Pour cela, le GETA a choisi de développer un système de gestion de bases de données lexicales, le système NADIA. Ce système est basé sur une approche interlingue. Comme unités interlingues, nous avons choisi d'utiliser les *acceptions* (sens généralement reconnus d'un mot). Le système NADIA fournit de nombreux outils pour la gestion d'une base lexicale multilingue. De plus, ce système laisse libre champ au linguiste quant aux structures linguistiques des entrées.

Cet article présente le projet NADIA. Après une étude générale de l'approche interlingue, nous présenterons le cadre dans lequel s'inscrit ce système, puis, nous donnerons une vue détaillée de l'approche par acceptions. Ensuite, nous étudierons le projet en présentant brièvement Parax, une étude de faisabilité, et l'architecture du système NADIA. Finalement, nous étudierons certains projets de bases de données lexicales. Nous verrons les choix faits par les responsables de ces projets et les problèmes auxquels ils ont été confrontés. Nous terminerons par une justification de nos choix, au vu de ces projets.

Approche par acceptions

Approche interlingue

L'approche interlingue utilise un langage artificiel intermédiaire (appelé *interlangue* et employé comme langage pivot) pour réaliser le lien entre les langues.

Les énoncés de toutes les langues considérées peuvent être représentés par cette

interlangue (indépendamment de la langue de l'énoncé). Aussi, une interlangue doit-elle avoir son propre lexique et son propre ensemble d'attributs et de relations.

Une interlangue doit être définie en référence à un certain ensemble de langues naturelles, à moins qu'un univers de référence fixe (ontologie) ne soit représenté de manière autonome par la machine.

Une interlangue consiste en deux parties distinctes : un lexique et un ensemble d'attributs et de relations.

Lexique interlingue

La première partie d'une interlangue est le lexique. Celui-ci doit être suffisamment complet pour représenter les différents sens des mots trouvés dans l'ensemble des langues considérées.

Ainsi, un lexique interlingue doit contenir au moins autant de sens de mots que chaque dictionnaire monolingue.

Comme une interlangue est définie pour 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.

Hélas, il n'y a pas nécessairement correspondance directe entre les sens des mots de différentes langues. Prenons l'exemple des mots français *fleuve* et *rivière* (dans leur sens concret le plus commun). Ces deux mots sont traduits en anglais par le mot *river* (dans son sens le plus commun). Les deux mots français ont deux sens différents[1]. L'anglais ne distingue pas ces deux sens. Un lien doit donc être établi entre ces sens dans le lexique interlingue. Par contre, 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 lieu d'être, puisque le mot japonais *kawa* recouvre le même sens que le mot anglais *river*.

Si l'interlangue est définie via un univers de référence fixe (ontologie), une description des différents sens des mots de chaque langue devra être donnée dans cet univers.

Dans ce cas, des sens équivalents de différentes langues devront avoir des descriptions identiques. De plus, des sens « proches » (comme *rivière* et *fleuve*) devront avoir des descriptions « proches ». Dans un contexte de traduction automatique, cette « distance » entre les descriptions devrait être automatiquement reconnue dans le cas où l'on n'a pas équivalence directe.

Attributs et relations interlingues

La seconde partie de l'interlingue est l'ensemble de ses attributs et relations. Cet en-

1. Une *rivière* est un cours d'eau se jetant dans un autre cours d'eau. Un *fleuve* est un grand cours d'eau se jetant dans la mer.

semble d'attributs et de relations doit être suffisamment complet pour permettre de coder les aspects linguistiques de toutes les langues considérées.

Cette partie n'est pas simple à définir, même si des études linguistiques fondamentales produisent de plus en plus de « microthéories » interlingues ou universelles (selon les termes de Nirenburg et Defrise (1990)) pour des phénomènes linguistiques, tels que l'aspect, le temps, la modalité, etc., qui, 20 ans plus tôt, semblaient ne pouvoir être décrits que par référence à une langue.

Projets utilisant l'approche interlingue

Certains projets ont adopté l'approche interlingue. Parmi ces projets, considérons l'américain KBMT-89 et le japonais EDR. Ces projets ont des buts différents. Le premier développe des dictionnaires pour une application particulière (un système de traduction basé sur la représentation des connaissances), alors que le second développe des bases lexicales pour différents systèmes de traduction automatique.

Le projet KBMT-89 (Gates *et al.* 1989 ; Nirenburg 1989) a défini et implémenté un système de traduction basé sur la connaissance du monde. Pour cela, un dictionnaire de 900 unités lexicales pour l'anglais et de 800 unités lexicales pour le japonais (représentant environ 1 500 concepts) a été développé.

Le projet EDR (EDR 1988, EDR 1990) vise au développement de ressources de grandes tailles. EDR a développé de grands dictionnaires d'environ 300 000 mots en anglais et 300 000 mots en japonais (200 000 mots de vocabulaire général et 100 000 en vocabulaire terminologique), ainsi qu'un dictionnaire de 400 000 concepts, un dictionnaire de cooccurrences (en anglais et japonais), ainsi que des dictionnaires bilingues anglais-japonais et japonais-anglais (300 000 entrées chacun).

Cadre du projet NADIA

Notre projet, NADIA *(Neutral Advanced Dictionaries by Interlingual Acceptions)* est un projet visant au développement d'un système de gestion de bases de données lexicales multilingues.

Nous visons quatre objectifs principaux :

- **Multilinguisme** : le système gère des bases de données multilingues. Aussi devrons-nous prendre en compte les différents systèmes d'écriture et les différentes procédures de tri.
- **Indépendance vis-à-vis des applications** : le système n'introduit pas de restriction sur les applications qui utiliseront les bases définies. Toute utilisation des bases est possible (pour traduction, correction, apprentissage des langues par l'homme), pourvu que l'information linguistique nécessaire soit présente.
- **Généricité** : les structures linguistiques utilisées par les bases seront définies par un linguiste, via un langage spécialisé.
- **Indépendance vis-à-vis des théories** : le système ne doit pas introduire de restrictions sur la théorie linguistique sous-jacente à une base. Il doit au

contraire permettre l'utilisation de nombreuses théories linguistiques qui mettent en œuvre de nombreux formalismes informatiques.

Nous avons choisi une approche interlingue car nous pensons ainsi apporter la meilleure solution au critère d'indépendance vis-à-vis des applications. Une telle approche assure la compatibilité avec une application interlingue. De plus, il est possible de générer des dictionnaires bilingues pour tout couple de langue présent dans la base.

Afin de réduire les problèmes de l'approche interlingue, nous avons choisi d'utiliser des acceptions comme unités interlingues. Une acception d'une langue est le sens particulier d'un mot, admis et reconnu par l'usage. Une acception, en tant qu'unité interlingue est une acception d'une des langues de la base.

Le projet ULTRA du CRL *(Computing Research Laboratory,* New Mexico State University) utilise une approche analogue (Farwell *et al.* 1992).

L'interlangue vue par NADIA

Afin de pouvoir fournir des outils utiles et puissants, le système de gestion de bases lexicales introduit des restrictions sur les bases elles-mêmes : les bases lexicales multilingues devront être basées sur l'approche interlingue par acceptions.

Architecture lexicale

L'architecture lexicale régit l'organisation des différents dictionnaires et leurs relations à l'intérieur d'une base lexicale.

Nous insistons sur le fait que cette architecture n'a pas d'influence sur le contenu linguistique des dictionnaires de la base, pas plus que sur le format que le linguiste adopte pour coder l'information.

Une base lexicale sous NADIA est composée de deux types de dictionnaires. Le premier type ne comprend qu'un dictionnaire : le dictionnaire interlingue. Le second type regroupe les dictionnaires monolingues.

Le dictionnaire interlingue contient les acceptions des différentes langues de la base.

Les dictionnaires monolingues contiennent l'information linguistique des différentes entrées lexicales. 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 entrées et acceptions. Un dictionnaire monolingue est généralement 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).

Lexique interlingue

L'utilisation des acceptions nous permet d'éviter les problèmes du raffinement des sens. Pour chaque langue, on peut choisir un dictionnaire existant comme référence. Les acceptions de la langue seront les acceptions trouvées dans ce dictionnaire de référence.

Le dictionnaire interlingue d'acceptions consiste en l'union ensembliste des acceptions des langues de la base. S'il y a correspondance directe entre deux acceptions de deux langues différentes, celles-ci sont confondues en une seule acception interlingue. S'il n'y a pas correspondance directe, les acceptions d'une langue sont conservées dans le dictionnaire interlingue.

La gestion d'un lexique interlingue est une tâche complexe. Aussi, cette gestion est assurée par la machine. Les modifications sur le lexique interlingue seront effectuées lorsque la machine détectera un problème.

Attributs et relations de l'interlangue

L'ensemble des attributs et des relations de l'interlangue n'est pas fixé. Cet ensemble est défini par un linguiste pour chaque base.

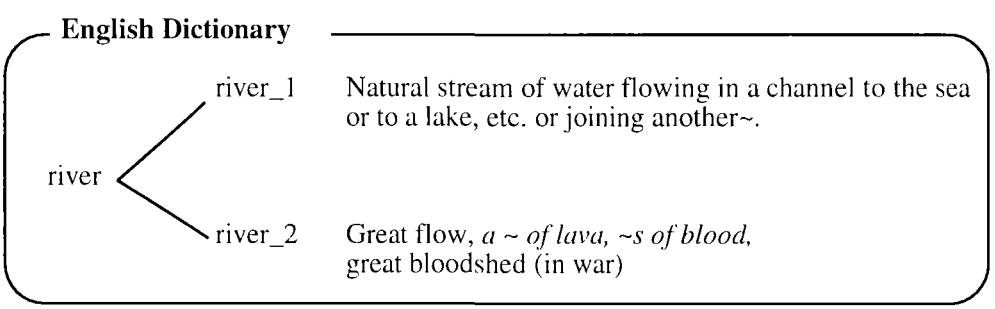
Par contre, afin de gérer les équivalences indirectes (comme dans l'exemple de *rivière, fleuve* et *river* vu plus haut), la relation entre acceptions « isa » (relation de sur à sous-acception) est prédéfinie. Ainsi, par cette relation, on code que les acceptions usuelles de *rivière* et de *fleuve* sont des sous-acceptions de l'acception usuelle de *river*.

Cette relation sera toujours définie dans le dictionnaire interlingue d'acceptions. Elle constitue l'ensemble minimal des relations et attributs interlingues.

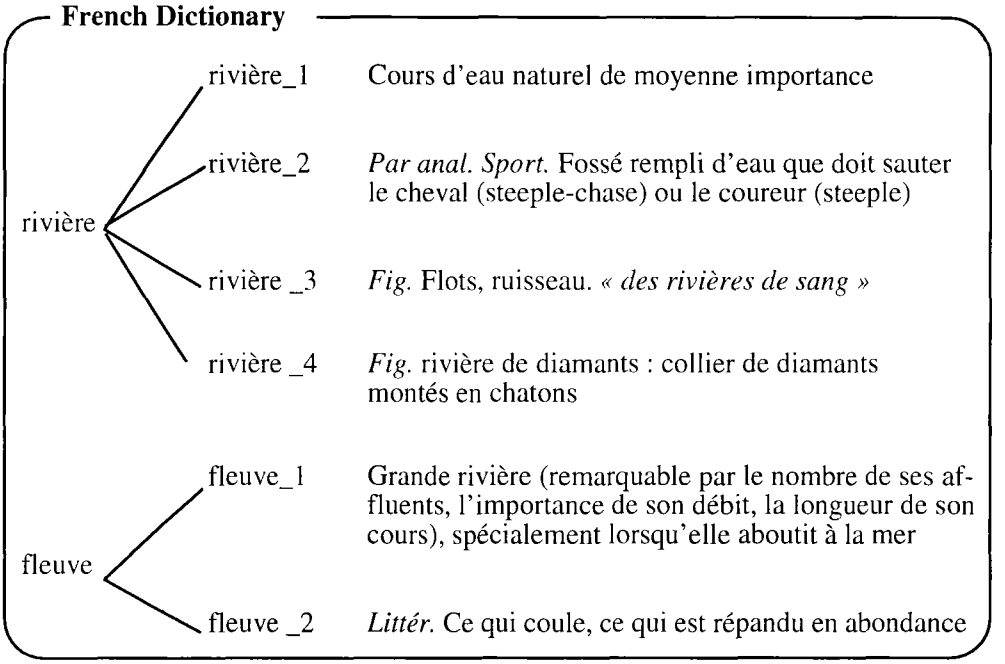
Un exemple

Reprenons l'exemple précédent plus en détail.

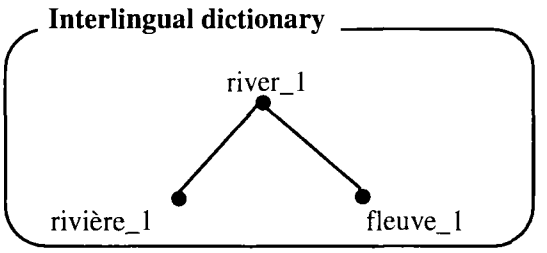
Dans la partie purement monolingue du dictionnaire anglais, l'entrée *river* a deux acceptions :



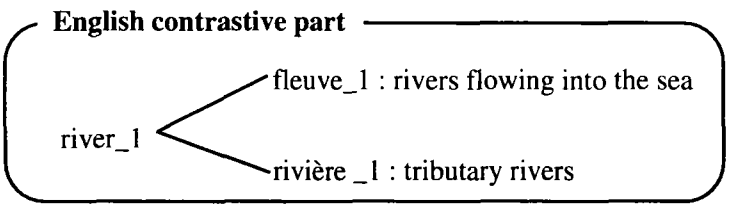
La partie purement monolingue du dictionnaire français contient les acceptions suivantes pour *rivière* et *fleuve*.



Comme le mot anglais *river* peut être traduit par *rivière* ou par *fleuve* suivant ce qu'il représente, le dictionnaire interlingue doit établir que les acceptions « rivière_1 » et « fleuve_1 » sont des sous-acceptions de l'acception « river_1 » .



La partie contrastive du dictionnaire anglais contient les informations permettant de distinguer les deux acceptions.



Le projet Nadia

Parax

Parax est une maquette implémentée sous HyperCard™ par Étienne Blanc pour étudier la faisabilité d'une telle approche. Cette maquette comporte pour l'instant cinq langues : français, anglais, allemand, russe et chinois.

La structure linguistique des dictionnaires monolingues est inspirée des dictionnaires du générateur de systèmes de traduction automatique ARIANE du GETA (Boitet *et al.* 1982 ; Boitet *et al.* 1985). Cette structure linguistique est fixe.

Cette maquette illustre le choix de l'interlangue par acceptions. L'utilisateur peut sélectionner un mot d'une des langues de la base. Puis, Parax lui propose une liste d'acceptions correspondant au mot choisi. Chaque acception est accompagnée d'une définition en français. Dans une première approche, les définitions sont stockées avec les acceptions dans la base interlingue (et non dans chacune des bases monolingues). Elles n'apparaissent donc que dans la langue de référence de la base : le français.

	fermer_1
fermer 1, a #fermer_fermeture$ CAT : vt. AUX : avoir.	**#fermer_fermeture$ °AL °AN °FR °RU** appliquer une partie mobile pour boucher un passage, une ouverture *(fermer la porte, les rideaux)* **#fermer_fermeture$organe °CH** fermer *(bouche, yeux...)* **#fermer_fermeture$norg °CH** fermer *(sauf bouche, yeux...)*
fermer 1, b #fermer_lieu CAT : vt. AUX : avoir.	**#fermer_lieu °AL °AN °FR °RU** priver de communication avec l'extérieur, par la mise en place d'un élément mobile ; interdire le passage *(fermer sa chambre, une valise, une route)*
fermer 1, c #fermer_replier$ CAT : vt. AUX : avoir.	**#fermer_replier$ °FR °RU** rapprocher, réunir (parties d'un organe, éléments d'un objet) de manière à ne pas laisser d'intervalle ou à replier vers l'intérieur, et par ext. mettre cet objet ou cet organe dans cette nouvelle position *(fermer le poing, un canif)* **#fermer_replier$poing °AL** comme fermer_replier$, mais s'applique particulièrement au poing (AL) **#fermer_replier$objet °AL** complément de fermer_replier$1 à fermer _ replier $

L'utilisateur peut ensuite choisir une des acceptions. Dans l'exemple fourni, on trouve un problème contrastif entre le français et l'allemand. Il existe deux acceptions allemandes différentes correspondant à l'acception française sélectionnée : « fermer_replier ». Si l'objet du verbe *fermer* est un poing, la traduction allemande sera *ballen*, sinon, elle sera *zuklappen.*

SOURCE : français	#fermer_replier$	p 133	CIBLE :
fermer 1, c #fermer_replier$ CAT : vt. AUX : avoir.	**#fermer_replier$ °FR °RU** rapprocher, réunir (parties d'un organe, éléments d'un objet) de manière à ne pas laisser d'intervalle ou à replier vers l'intérieur, et par ext. mettre cet objet ou cet organe dans cette nouvelle position *(fermer le poing, un canif)* **#fermer_replier$poing °AL** comme fermer_replier$, mais s'applique particulièrement au poing (AL) **#fermer_replier$objet °AL** complément de fermer_replier$1 à fermer_replier$		

Si l'utilisateur veut un équivalent russe, il n'a qu'à choisir le russe comme langue cible.

SOURCE : français	#fermer_replier$	p 133	CIBLE : russe
fermer 1, c #fermer_replier$ CAT : vt. AUX : avoir.	**#fermer_replier$ °FR °RU** rapprocher, réunir (parties d'un organe, éléments d'un objet) de manière à ne pas laisser d'intervalle ou à replier vers l'intérieur, et par ext. mettre cet objet ou cet organe dans cette nouvelle position *(fermer le poing, un canif)* **#fermer_replier$poing °AL** comme fermer_replier$, mais s'applique particulièrement au poing (AL) **#fermer_replier$objet °AL** complément de fermer_replier$1 à fermer_replier$		закрыть #fermer_replier $

Par contre, s'il désire la traduction allemande, il lui faudra sélectionner la sous-acception voulue, puis la langue cible.

SOURCE : français	#fermer_replier$	p 133	CIBLE : allemand
fermer 1, c #fermer_replier$ CAT : vt. AUX : avoir.	**#fermer_replier$ °FR °RU** rapprocher, réunir (parties d'un organe, éléments d'un objet) de manière à ne pas laisser d'intervalle ou à replier vers l'intérieur, et par ext. mettre cet objet ou cet organe dans cette nouvelle position *(fermer le poing, un canif)* **#fermer_replier$poing °AL** comme fermer_replier$, mais s'applique particulièrement au poing (AL) **#fermer_replier$objet °AL** complément de fermer_replier$1 à fermer_replier$		ballen #fermer_replier$poing

L'utilisateur peut ainsi voir le terme français et son équivalent allemand (pour l'acception donnée).

NADIA : Architecture logicielle

La maquette Parax n'est qu'une illustration de l'approche interlingue que nous avons choisie. Elle n'autorise 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.

Le projet NADIA est la seconde étape dans le développement d'un système de gestion de bases lexicales. Ce projet vise à la construction d'un prototype aussi complet que possible d'un tel système. Ce prototype inclut tous les outils prévus pour la gestion de bases lexicales. Il est construit sans souci de performance en termes de nombre d'unités lexicales.

Ce prototype est développé sur Macintosh™, avec MCL *(Macintosh Common Lisp)* et CLOS *(Common Lisp Object System)*. Il utilise les techniques de programmation par objets.

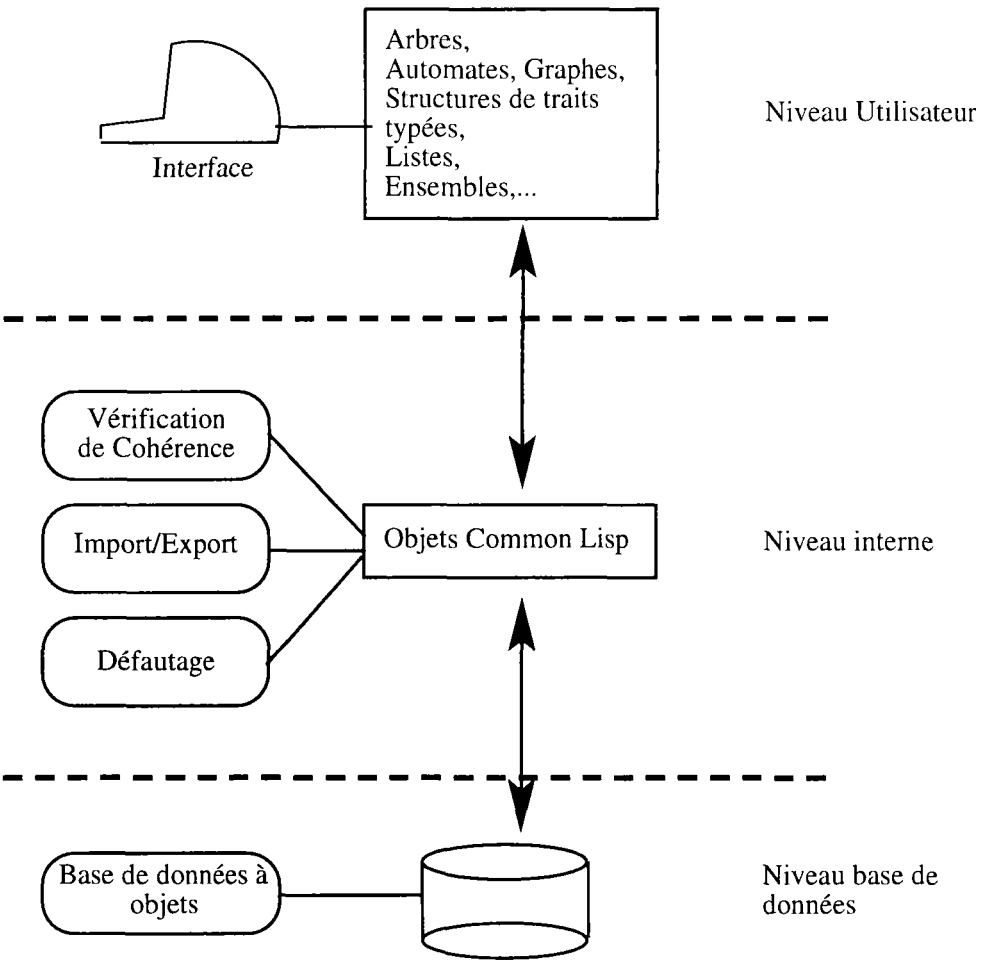
L'architecture logicielle de NADIA se compose de 3 parties :

- Le *niveau base de données* est le niveau le plus bas. C'est à ce niveau que les objets seront archivés ou retrouvés. Le prototype utilise une base de données à objets du domaine public écrite en MCL : WOOD[2]. Ce niveau est complètement transparent aux utilisateurs du système (lexicographes, linguistes,...).
- Le *niveau interne* est le niveau où travaillent les différents outils linguistiques.

2. WOOD signifie *William's Object Oriented Database*. Il s'agit d'un programme permettant de manipuler des objets persistants en MCL. Il est écrit par Bill St. Clair (bill@cambridge.apple.com).

Les unités linguistiques sont représentées comme des objets de *Common Lisp*. Ce niveau est transparent à l'utilisateur.

- Le *niveau utilisateur* est le niveau abstrait où travaille l'utilisateur. À ce niveau, l'utilisateur manipule des objets linguistiques (arbres, structures de traits typées, listes, automates, graphes, ensembles,...). Les différents outils communiqueront à ce niveau.



Le fonctionnement de cette architecture est basé sur l'aller et retour entre les niveaux utilisateur, interne et base de données.

La définition de la structure des entrées d'un dictionnaire est faite au niveau utilisateur en termes de structures linguistiques usuelles (arbres, structures de traits, graphes,...). Cette définition est ensuite traduite en termes d'objets *Common Lisp*.

La définition des règles de cohérence et d'intégrité est faite par le linguiste au niveau utilisateur. Ces règles sont « compilées » et utilisées au niveau interne. Elle sont stockées dans la base.

Les requêtes d'interrogation de la base sont exprimées au niveau utilisateur. Elles sont traduites et évaluées au niveau interne. Le résultat est présenté au niveau utilisateur.

L'utilisateur peut exporter et importer des entrées ou parties d'entrées. La définition des règles d'import et d'export est faite au niveau utilisateur, les règles sont compilées et effectivement appliquées au niveau interne. Le résultat des procédures d'export est stocké sous un format SGML *(Standard Generalised Markup Language)* reflétant les structures linguistiques (les conventions de TEI *(Text Encoding Initiative)* seront suivies autant que possible).

Gestion des dictionnaires

On trouve deux sortes de dictionnaires dans une base lexicale multilingue NADIA :

1. *Les dictionnaires monolingues* sont divisés en deux parties :
   - une partie purement monolingue qui contient les acceptions de la langue et leurs informations linguistiques associées ;
   - une partie contrastive qui contient les acceptions existant dans d'autres langues, mais pas dans la langue du dictionnaire, ainsi que les informations associées.
2. *Le dictionnaire interlingue* contient les acceptions interlingues et leurs relations.

Le linguiste définit les structures linguistiques des entrées de chaque dictionnaire monolingue via un langage spécialisé. Il est aussi possible de définir des contraintes de cohérence et d'intégrité sur ces structures, ainsi que des règles de valeurs par défaut. Le linguiste gère les informations des dictionnaires monolingues, à l'aide d'outils de NADIA (éditeur, défauteur, vérificateur de cohérence,...)

Le dictionnaire interlingue est difficile à gérer. Une acception interlingue est créée lorsqu'une nouvelle acception apparaît dans une langue. Le lexique interlingue doit fournir des liens lexicaux entre les différentes langues. Pour cela, les acceptions équivalentes de deux langues différentes doivent être réunies en une seule acception interlingue. Aussi, lorsqu'il ajoute une nouvelle acception dans un dictionnaire, le lexicographe doit vérifier si une acception interlingue équivalente existe dans la base interlingue. Si oui, il va lier cette nouvelle acception à l'acception interlingue. Si non, il doit créer une nouvelle acception interlingue. Une telle gestion de la base interlingue suppose que :

- le lexique interlingue fournisse suffisamment d'informations pour que le lexicographe puisse vérifier l'existence d'une acception interlingue. Cette information doit définir chaque acception de manière non ambiguë ;
- cette information soit comprise par tous les lexicographes. Elle doit donc être fournie dans une langue commune ;

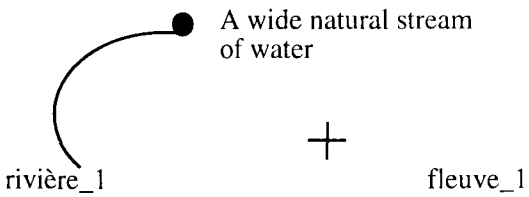
- le lexicographe qui crée une nouvelle acception fournisse cette information dans la langue commune (qui n'est pas nécessairement sa langue maternelle).

Pour ces raisons, nous avons choisi de confier la gestion de la base interlingue au système. Celui-ci crée une nouvelle acception dans le dictionnaire interlingue lorsqu'une nouvelle acception apparaît dans un dictionnaire monolingue. Il lie un nouveau terme avec une acception interlingue déjà existante lorsque cela est possible.

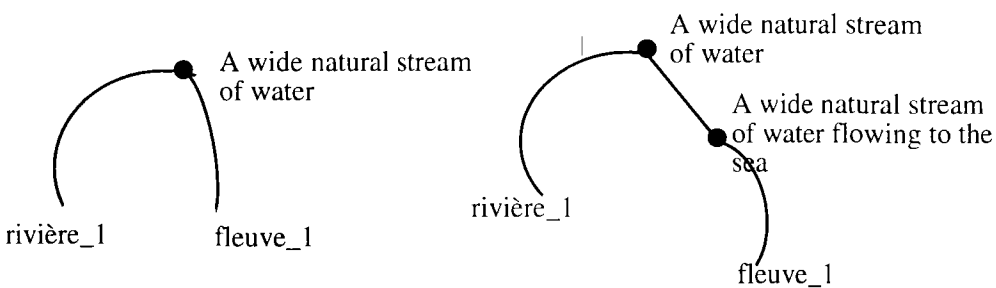
La détermination de l'existence d'une acception interlingue ou du besoin d'en créer une nouvelle se fait par un « effet de bord ». On demande au lexicographe de donner l'équivalent du terme en cours dans une des langues qu'il connaît le mieux (parmi celles de la base). Le système sait ainsi si l'acception interlingue correspondante existe ou non, et crée une nouvelle acception si nécessaire.

Pour illustrer les principes de la gestion du dictionnaire interlingue, prenons l'exemple de l'indexage des deux mots français *rivière* et *fleuve.* Supposons que la base lexicale contient déjà un dictionnaire anglais.

Le lexicographe indexe le mot français *rivière.* Il sélectionne l'acception qui convient parmi les différentes acceptions de l'équivalent anglais : *river* et la lie au nouveau mot français. Quand il indexera le mot *fleuve,* il retrouvera la même acception de *river.*

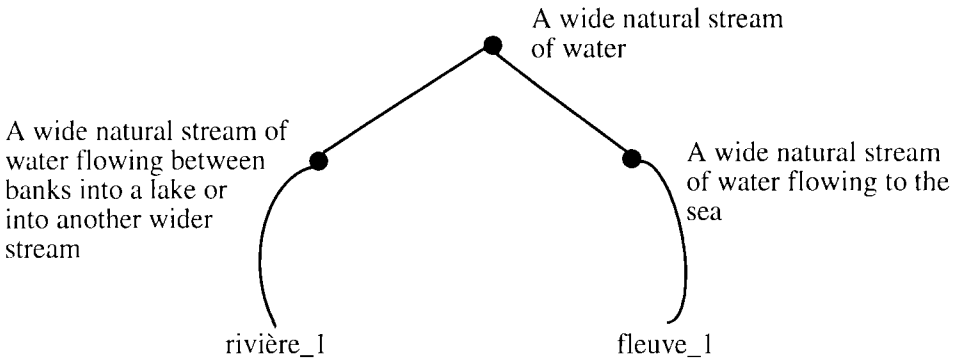


Devant ce problème, le lexicographe pourrait adopter plusieurs choix : lier cette acception au mot *fleuve* ou créer une nouvelle acception pour *fleuve* et la lier à l'acception de *river.*



Le premier choix est incorrect, puisque l'acception est déjà liée à *rivière* et que *rivière* n'est pas synonyme de *fleuve.* Le second l'est aussi, puisque *fleuve* ne représente pas un sous-sens de *rivière.*

Le choix correct pour résoudre ce problème est de modifier le lien entre *rivière* et l'acception originale (celle qui correspond à *river*) en créant deux nouvelles acceptions (une pour *rivière*, l'autre pour *fleuve*). Ces deux nouvelles acceptions sont des sous-acceptions de l'acception originale.



Afin d'assurer une cohérence dans la gestion du dictionnaire monolingue, et pour alléger le travail du lexicographe, le système prend en charge cette gestion en détectant automatiquement les différents problèmes qui peuvent se poser et en proposant des solutions au lexicographe.

Justification des choix

Analyse de projets existants

KBMT-89

KBMT-89 *(Knowledge Based Machine Translation)* était un projet de recherche du *Center for Machine Translation* de l'université Carnegie Mellon. Le but de ce projet est de construire un système de traduction automatique utilisant une approche basée sur la connaissance. Le système KBMT-89 utilise une représentation de la connaissance du domaine en tant qu'interlangue.

La connaissance acquise pour ce système inclut une ontologie (environ 1 500 concepts). Les lexiques contiennent environ 800 unités lexicales japonaises et 900 unités anglaises. On trouvera une description plus détaillée dans (Meyer *et al.* 1990 ; Nirenburg 1989).

Les dictionnaires de KBMT-89 sont composés de :

- une base des connaissances (ontologie) ;
- des lexiques monolingues (japonais et anglais).

La base des connaissances contient un ensemble de concepts, classés en une hiérarchie. Ces concepts servent d'interlangue.

Les lexiques contiennent un ensemble de « super-entrées » (lemmes) décomposées en « entrées » (sens). À chacun de ces sens sont associées des informations linguistiques (catégorie, orthographe, phonologie, morphologie, traits syntaxiques, structures syntaxiques, sémantiques, relations lexicales et pragmatiques).

Chaque sens d'un mot est généralement relié à un concept de la base de connaissance. Un sens peut être relié à une structure interlingue qui n'est pas reflétée par un concept de l'ontologie (une attitude ou une relation).

EDR

EDR *(Electronic Dictionary Research)* est un projet japonais visant à la construction d'une base lexicale de grande taille pour différents systèmes de traduction automatique. EDR a développé des dictionnaires de mots, des dictionnaires de concepts, des dictionnaires bilingues et des dictionnaires de cooccurrences pour l'anglais et le japonais.

Le dictionnaire de concepts de EDR contient une description des concepts. Cette description se fait au travers de relations entre concepts. EDR utilise 32 types de relations et cinq types d'attributs pour cette description. Les concepts sont classifiés dans une hiérarchie.

Le dictionnaire de concepts est utilisé comme un dictionnaire interlingue. Le dictionnaire de mots fait le lien entre mot et concept, et contient les informations linguistiques associées à chaque mot.

Multilex

Le projet européen Multilex (projet DG XIII - ESPRIT) vise à définir des standards pour la construction de bases lexicales multilingues. Les langues considérées sont celles de la Communauté européenne.

Multilex utilise une approche par transfert. Une base lexicale Multilex comporte deux sortes de dictionnaires : des dictionnaires monolingues et des dictionnaires bilingues.

Les dictionnaires monolingues contiennent des unités lexicales (sens) ainsi que l'information linguistique associée. Un langage a été développé pour que le linguiste puisse définir les structures utilisées. Ces structures doivent être codées sous forme de structures de traits typées.

Les dictionnaires bilingues sont unidirectionnels. Chaque unité bilingue est composée d'une unité lexicale source, d'une unité lexicale cible, d'une condition d'application et d'une règle de transformation lorsque nécessaire.

Problèmes rencontrés

Il est intéressant d'analyser les principaux problèmes rencontrés par les projets de

bases lexicales multilingues. Tous rencontrent des problèmes liés au codage des informations linguistiques, et les projets utilisant un vocabulaire interlingue butent sur le raffinement des concepts. Enfin, nous parlerons des problèmes introduits par la représentation des connaissances, puis de ceux introduits par une classification des concepts.

Représentation des informations linguistiques

Le projet EDR représente les informations linguistiques sous forme d'attributs. Les attributs possibles sont définis *a priori* et ne peuvent être modifiés. L'information linguistique n'est pas très détaillée. Mais les informations présentes sont suffisantes pour être utilisées par des systèmes de traduction automatique.

Le projet Multilex est plus souple puisqu'il permet au linguiste de définir la structure linguistique des dictionnaires. Par contre, celui-ci est obligé d'utiliser des structures de traits typées pour coder ses informations lexicales. Bien adaptées pour certains problèmes, les structures de traits typées le sont moins pour coder certaines structures linguistiques telles que les arbres ou les automates.

Raffinement des concepts

Pour utiliser des concepts dans l'interlangue, on doit répondre à la question : « Qu'est-ce qu'un concept ? ». Prenons par exemple les trois phrases suivantes :

- Un éléphant apparaît.
- Un éléphant est un animal intelligent.
- L'éléphant est une espèce en danger.

Pour *éléphant,* on doit savoir si l'on a affaire à trois concepts différents ou à trois réalisations différentes d'un même concept.

EDR considère que les trois phrases contiennent trois réalisations différentes d'un même concept (la première phrase parle d'un éléphant en tant qu'individu, la seconde en tant qu'éléphant typique et la troisième, en tant qu'espèce animale).

KBMT-89 a défini quatre critères pour raffiner les concepts. L'un de ces critères déclare que si une unité lexicale a deux ensembles d'attributs grammaticaux incompatibles, alors deux unités lexicales doivent être créées. Les attributs grammaticaux peuvent être morphologiques (exemple : nom discret ou non), syntaxiques ou lexicaux (exemple : collocations). Mais ce critère n'est pas absolu : il dépend de la taille et de la systématicité des différences entre les deux ensembles.

Dans l'exemple précité, il n'est pas aisé de voir si l'on a affaire à un ou plusieurs concepts. Dans la première phrase, le nom *éléphant* est un nom discret (il est possible de dire *Trois éléphants apparaissent).* Dans la seconde et la troisième phrase, le nom *éléphant* relève du non-discret (exemple : dire *Trois éléphants sont des animaux intelligents* n'aurait pas la même signification). De plus dans la troisième phrase, on ne peut pas utiliser l'article indéfini.

On doit faire un choix entre créer trois concepts ou un seul. Lorsque le choix est fait (quel qu'il soit), le même choix doit être fait à chaque fois qu'un cas semblable se présente. Lorsque l'on construit une base de taille réelle, plusieurs lexicographes doivent intervenir. Assurer la cohérence des choix devient alors un problème méthodologique complexe.

Problèmes introduits par la représentation des connaissances

EDR et KBMT-89 ont choisi d'utiliser une représentation des connaissances en tant qu'interlangue.

Ce choix introduit des problèmes méthodologiques lorsque l'on veut construire des bases de taille réelle.

Premièrement, il est coûteux et difficile de décrire les concepts pour une base à grande échelle (pour EDR, qui a décrit avec succès 400 000 concepts, il a fallu 1 200 hommes-années). Pour être réellement envisageable, une telle description devra se faire par une extraction automatique ou semi-automatique des concepts. EDR a largement utilisé une telle approche (à partir d'un grand corpus).

Lors de cette extraction des connaissances, EDR a utilisé un grand corpus de textes. Pour des raisons de rapidité et de complexité lors de l'acquisition, EDR a dû faire des généralisations hâtives (H. Suzuki, communication personnelle). Prenons par exemple la phrase *Un éléphant mange une pomme*. Cette phrase dit qu'un certain éléphant est en train de manger une certaine pomme. EDR en extrait deux relations pour le dictionnaire de concepts :

- Tout éléphant peut être agent du verbe *manger*,
- Toute pomme peut être l'objet de l'action *manger*.

Problèmes introduits par une classification des concepts

EDR et KBMT-89 réduisent le nombre de relations entre concepts en les factorisant. Cette factorisation se fait à l'aide d'une hiérarchie et d'un mécanisme d'héritage.

Ainsi, par exemple, dans la hiérarchie de EDR, le fait qu'un oiseau peut voler est codé. Le mécanisme d'héritage permet donc de savoir qu'un moineau, qu'une alouette, qu'une hirondelle,... peut voler.

Par contre cette classification des concepts peut poser des problèmes théoriques et méthodologiques.

Il faut tout d'abord gérer les exceptions. Comment doit-on représenter le fait qu'une autruche est un oiseau et qu'il ne peut voler ? EDR a choisi de déclarer explicitement qu'une autruche ne peut voler via une relation négative (H. Suzuki, communication personnelle). Cette assertion remplace l'assertion héritée. Hélas, de telles relations négatives doivent être ajoutées avec soin et ne peuvent pas être insérées automatiquement puisqu'elles peuvent introduire des incohérences dans la base.

La modification d'une telle hiérarchie est très délicate, puisque tous les sous-concepts du concept modifié seront affectés.

L'ajout d'une relation à un concept dans la hiérarchie peut s'avérer délicat si celui-ci a de nombreux sous-concepts. En effet, cette relation sera héritée par l'ensemble des sous-concepts. Il faudra donc vérifier qu'il n'y a pas de nouvelles exceptions parmi les sous-concepts.

Justification des choix

Les choix que nous avons faits pour l'organisation générale du projet NADIA sont justifiés tant par notre désir d'assurer à terme une certaine compatibilité (au moins en termes d'échange de données) avec les projets précités que par les problèmes qu'ils ont rencontrés.

Multilex utilise une approche par transfert, alors que EDR et KBMT-89 ont choisi une approche interlingue. Afin de pouvoir être compatible avec chacune de ces approches, nous avons choisi l'approche interlingue. Avec une telle approche, il est possible de générer des dictionnaires de transfert, et donc de disposer de dictionnaires compatibles avec ceux de Multilex.

Le projet EDR utilise une structure linguistique relativement figée. Multilex est plus souple, mais ne permet d'utiliser que des structures de traits typées. Pensant que la souplesse est un atout important, nous avons choisi de donner au linguiste la possibilité de définir ses structures linguistiques. Pour augmenter encore cette souplesse, nous n'imposons pas une structure informatique de base pour coder les structures linguistiques. Le système NADIA permet d'utiliser la plupart des structures informatiques les plus utilisées à l'heure actuelle en TALN. Il est aussi possible de mélanger ces différentes structures dans une même entrée, afin de disposer d'une représentation adaptée à chaque problème linguistique. C'est ainsi que nous garantissons une indépendance du système vis-à-vis de la théorie linguistique choisie.

Les deux projets interlingues étudiés utilisent une approche par représentation des connaissances. Nous avons vu que cette approche pose certains problèmes théoriques et méthodologiques. Pour tenter d'apporter une solution à certains de ces problèmes, nous avons choisi l'approche par acceptions interlingues, qui permet d'éviter le problème du raffinement des unités interlingues (ce raffinement devient systématique, puisqu'on utilise des dictionnaires existants comme référence). Elle nous permet aussi de ne pas représenter les connaissances du monde et de réduire ainsi considérablement le coût de la construction d'une base.

Le projet EDR utilise une classification de concepts pour réduire le nombre de relations de description. Comme nous n'utilisons pas de description de concepts, cette classification devient inutile. Nous n'utilisons qu'une sorte de relation entre acceptions : une relation de suracception à sous-acception. Celle-ci nous permet de coder les phénomènes contrastifs. Il n'est cependant pas question de développer une hiérarchie complète d'acceptions par le biais de cette relation. Celle-ci ne sera utilisée que localement, en cas de problème contrastif.

Conclusion

Nous avons présenté ici les grandes lignes de notre projet NADIA, qui vise à la construction d'un système de gestion de bases de données lexicales multilingues permettant aux utilisateurs de définir des bases multilingues, indépendamment des applications qui utiliseront les données.

Le projet NADIA utilise une approche interlingue originale : l'interlingue par acceptions. Cette approche permet de s'affranchir des problèmes de représentation de connaissances que l'on rencontre souvent dans les projets interlingues.

En donnant la possibilité au linguiste de choisir la (ou les) structure(s) linguistiques et informatiques qu'il désire, nous garantissons une certaine dépendance par rapport à la théorie linguistique sous-jacente de chacune des bases. Nous effectuons ainsi un nouveau pas vers le partage des données linguistiques en permettant à différentes théories linguistiques de cohabiter sur une seule et même plate-forme.

Le prototype de NADIA décrit ici est en cours de développement. Après ce prototype, plusieurs voies s'offriront à nous. Il nous sera possible d'améliorer les différents outils, et notamment les outils d'import/export qui permettront encore une fois un partage de données linguistiques entre différents systèmes.

Une nouvelle voie de recherche se dessine dès à présent. Nous faisons l'analogie entre un dictionnaire et un document structuré. De la même manière qu'un document est une suite de chapitres (ayant un titre) composés de parties contenant elles-mêmes différents paragraphes, un dictionnaire est une suite d'articles (ayant une forme d'entrée) composés de différents sens, contenant eux-mêmes différentes informations linguistiques.

Une telle analogie nous permet d'envisager, comme pour un document structuré, de définir différentes vues d'une base lexicale. Ainsi, un lexique quadrilingue en colonnes pourra être une vue d'une base lexicale quadrilingue. Un fichier SGML pourra aussi être une vue d'une base lexicale. Cela nous permettra de confondre les problèmes d'interface et d'import/export. Nous pourrons ainsi étudier comment définir différentes théories linguistiques en tant que différentes vues d'une seule et même structure.

Nous espérons ainsi faciliter de plus en plus, non pas seulement le partage de données linguistiques, mais aussi la communication entre différentes « écoles » linguistiques, qui pourraient mettre en commun, sur une seule base, les différents aspects qui les intéressent, dans leurs codages préférés.

Références

BOITET, C., GUILLAUME, P. et M. QUEZEL-AMBRUNAZ (1982) : « ARIANE-78: an Integrated Environment for Automatic Translation and Human Revision », *COLING-82*, juillet 1982, pp. 19-27.

BOITET, C., GUILLAUME, P. et M. QUEZEL-AMBRUNAZ (1985) : « A Case Study

in Software Evolution: from ARIANE-78.4 to ARIANE-85», *Theoretical and Methodological Issues in Machine Translation of Natural Languages*, 14-16 août 1985, vol. 1/1, pp. 27-58.

EDR (1988) : *Electronic Dictionary Project*, Japan Electronic Dictionary Research Institute Ltd, novembre 1988.

EDR (1990) : *EDR Technical Reports, an Overview of the Electronic Dictionaries*, EDR, Japan Electronic Dictionaries Research Institute Ltd, Technical reports n° TR-024, TR-025, TR-026, TR-027, TR-029, 176 p.

FARWELL, D., GUTHRIE, L. et Y. WILKS (1992) : « The Automatic Creation of Lexical Entries for a Multilingual MT System », *Proceedings of the 14th International Conference on Computational Linguistics*, COLING-92, Nantes, 20-28 juillet 1992, vol. 2/4, pp. 532-538.

GATES, D. *et al.* (1989) : « Lexicons », *Machine Translation*, vol. 4-1, pp. 67-112.

MEYER, I., ONYSHKEVYCH, B. et L. CARLSON (1990) : *Lexicographic Principles and Design for Knowledge-Based Machine Translation*, Carnegie Mellon University, Technical Report n° CMU-CMT-90-118, 13 août 1990, 66 p.

NIRENBURG, S. (1989) : « Knowledge-based Machine Translation », *Machine Translation*, vol. 4-1, pp. 5-24.

NIRENBURG, S. et C. DEFRISE (1990) : « Lexical and Conceptual Structure for Knowledge-based Machine Translation », *ROCLING III*, 20-22 Août 1990, vol. 1-1, pp. 105-130.

7

À propos de la traduction (automatique) de *faire* en anglais

Laurence DANLOS*

TALANA, Université Paris VII, Paris, France

• ***Abstract*** •

The verb faire *is widely used in written French, and more than widely in spoken French. It translates in English in various ways according to the noun it introduces. However, it is out of the question to compute its translation by means of lexical context sensitive rules,* e.g. faire (~ promenade) —-> take, *because of the huge number of such ad hoc rules. Our solution consists in isolating some classes of nouns and coding the right lexical information under the noun. In addition to this solution for processive nouns as* promenade *(where* faire *is a support verb), we will put forward this solution for other classes of nouns.*

Le verbe *faire* est un verbe courant du français qui entre dans de nombreuses constructions et reçoit des traductions anglaises variées. Parmi ces constructions, on observe les constructions à verbe support (Giry-Schneider 1978 ; 1987) telles que :

Jean a fait une promenade (--> *John took a walk*)
Jean fait des complexes (--> *John has complexes*)
Jean a fait des progrès (--> *John made progresses*)
Jean a fait une farce à Mary (--> *John played a trick on Mary*)

Une solution pour traiter automatiquement en analyse et en traduction les constructions à verbe support (avec le verbe *faire* ou un autre verbe support comme *prendre,*

* Je tiens à remercier Lee Humphreys et Louisa Sadler pour leurs informations sur la traduction anglaise des phrases avec *faire*, et Pollet Samvelian pour la relecture de cet article.

avoir ou *donner*) est décrite dans (Danlos 1992 ; Danlos et Samvelian 1992). Cette solution repose sur des informations lexicales codées sous des noms comme *promenade, complexe, progrès* ou *farce*, et évite ainsi les règles bilingues sensibles au contexte lexical comme *faire (- promenade) --> take*. Gardant à l'esprit qu'il faut à tout prix éviter les règles bilingues sensibles au contexte lexical pour calculer la traduction de *faire*, vu le coût de ces règles qui peuvent se compter par milliers, nous nous consacrerons dans cet article à la construction suivante :

(A) N_0 (+ humain) faire N_1 (+ artefact)[1]
= Jean a fait un gâteau
= Jean a fait un livre sur ce sujet

dans laquelle N_1 désigne un « artefact », c'est-à-dire un objet créé par un humain. Un artefact peut désigner soit un objet concret comme *gâteau, robe, voiture* ou *maison*, soit un objet à caractère plus abstrait mais qui peut avoir une réalisation matérielle comme *livre, symphonie, liste* ou *carte*. La classe des artefacts s'oppose aux objets concrets qui existent naturellement comme *fleur, rocher* ou *soleil*, et aux noms purement abstraits comme *promenade, complexe, progrès* ou *farce*. Elle est souvent mentionnée dans la littérature bien qu'à notre sens elle ne soit pas clairement définie : ainsi nous hésitons sur le statut de noms comme *préface, conférence* ou *discours*, sont-ils des artefacts ou pas ? Néanmoins, ces hésitations sur les limites de cette classe, à laquelle nous avons recours principalement à titre de commodité, n'affectent pas notre propos, comme cela deviendra évident par la suite.

Le verbe *faire* dans la construction (A) peut induire différents types de relations entre N_0 et N_1 que nous allons examiner successivement.

N_0 est « l'auteur » de N_1

Le nom N_1, désignant un artefact, est créé par un « auteur » et la construction (A) est fréquemment utilisée pour indiquer que le sujet humain N_0 est l'auteur de N_1. Dans ce cas là, le verbe *faire* est souvent employé à la place d'un (ou de plusieurs) verbe(s) « de création » plus spécifique(s)[2] :

(1) Jean a (préparé + fait) un gâteau
(2) Jean a (confectionné + fait) une robe
(3) Jean a (fabriqué + fait) une voiture
(4) Jean a (construit + fait) une maison
(5a) Jean a (écrit + fait) un livre sur ce sujet
(6) Jean a (composé + écrit + fait) quatre symphonies
(7) Jean a (dressé + fait) une liste de termes
(8) Jean a (dessiné + fait) une carte du Mali
(9) Jean a (peint + crayonné + gouaché + fait) un portrait de Marie

1. Nous ne qualifions pas *faire* de verbe support dans la construction (A), bien que pour Giry-Schneider (1987) certains exemples de structure (A) mettent en jeu le verbe support *faire*. Sur le statut de *faire* et la notion de verbe support, voir Danlos et Samvelian (à paraître).
2. Ces verbes appartiennent souvent à la table 32A de Boons *et al.* (1976).

bien que l'existence d'un verbe de création spécifique ne soit pas systématique :

(10) Jean a fait pipi[3]
(11) Jean a (pris + fait) une photo de Marie[4]

Nous ne nous aventurerons pas à poser que tous les noms désignant un artefact entrent dans la construction (A) avec l'interprétation où N_0 est l'auteur de N_1. D'abord, rappelons que la notion d'artefact reste mal cernée à nos yeux, ce qui nous empêche de lancer une affirmation péremptoire à son sujet. Ensuite, certains noms désignent clairement un artefact mais ne permettent guère que la construction (A) ait l'interprétation où N_0 est l'auteur de N_1 :

(12) Jean a fait des mots croisés
(13) Jean a fait son lit

Dans (12), l'interprétation où Jean est l'auteur des mots croisés ne peut être obtenue que dans un contexte fortement marqué ; de même, il est difficile d'interpréter (13) dans le sens où Jean a fabriqué son lit. Les exemples comme (12) et (13) seront traités plus loin.

Si la notion « N_0 être l'auteur de N_1 » semble claire à l'intuition, il existe en fait des cas où elle ne l'est pas. Considérons un nom comme *livre. A priori*, l'auteur du livre est la personne qui l'a écrit. Mais si on considère que le fabricant d'une voiture est l'auteur de la voiture, alors l'imprimeur d'un livre en est aussi l'auteur, et pourquoi pas aussi son éditeur. Le verbe *faire* peut remplacer le verbe spécifique dans ces trois sens de *auteur d'un livre :*

(5a) Jean a (écrit + fait) un livre sur ce sujet
(5b) Jean a (imprimé + fait) un livre sur ce sujet
(5c) Jean a (publié + fait) un livre sur ce sujet

ce qui confère trois interprétations à la phrase

(5) Jean a fait un livre sur ce sujet

même si l'interprétation (a) est la plus naturelle. La phrase (5) a aussi l'interprétation

Jean a (étudié + fait) un livre sur ce sujet

où Jean n'est pas l'auteur du livre (voir la section sur les interprétations diverses de la construction (A)) et éventuellement l'interprétation :

Jean a rendu un livre sur ce sujet célèbre

Mais les interprétations de (5) s'arrêtent là. Autrement dit, la phrase (5) ne peut (pratiquement) jamais avoir le sens d'une des phrases suivantes :

3. Est-ce un artefact ?
4. Nous ne considérons pas que le verbe *prendre*, qui est fréquemment utilisé comme verbe support, soit un verbe spécifique.

Jean a (lu + acheté + emprunté + colorié + déchiré) un livre sur ce sujet[5]

On retiendra donc que la construction (A) peut avoir différents sens – identifiables par le verbe spécifique que *faire* remplace – mais que ces différents sens sont en nombre fini et donc listables. Ainsi, on peut compter cinq sens pour la phrase (5) dont trois relèvent de l'interprétation « N_0 être l'auteur de N_1 ».

Lorque *faire* est un pro-verbe d'un verbe de création plus spécifique, il partage des propriétés syntaxiques avec celui-ci. Ainsi *faire* comme *écrire* dans (5) permet la double analyse (Giry-Schneider 1978) :

C'est un livre sur ce sujet que Jean a (écrit + fait)
C'est sur ce sujet que Jean a (écrit + fait) un livre

tandis que *faire* comme *dresser* dans (7) ne permet pas cette double analyse :

C'est une liste de termes que Jean a (dressée + faite)
*C'est de termes que Jean a (dressé + fait) une liste

On soulignera cependant la disparité suivante : dans certains cas, le verbe spécifique que *faire* remplace peut admettre l'omission de l'objet direct, par exemple : *Jean a écrit sur ce sujet,* ce qui n'est jamais possible avec *faire*, par exemple : **Jean a fait sur ce sujet.* Il peut exister d'autres différences entre la construction avec *faire* et celle avec un verbe spécifique de création (par exemple : *écrire* se construit avec une complétive, ce qui n'est pas le cas de *faire : Jean a (écrit + *fait) que P sur ce sujet*), mais ces différences ne nous concernent pas ici où nous nous concentrons sur la construction (A).

Tournons-nous vers la traduction anglaise de (A) dans l'interprétation où N_0 est l'auteur de N_1. Pour des artefacts désignant des objets concrets, il semble que *faire* se traduise régulièrement par *make* (qui peut éventuellement lui aussi remplacer un verbe plus spécifique) :

(1') *John (baked + made) a cake*
(2') *John (? + made) a dress*
(3') *John (manufactured + constructed + made) a car*
(4') *John (built + made) a house*
(5'b) *John (printed + made) a book on this topic*

On notera d'ailleurs que ces objets concrets peuvent être vendus avec l'étiquette *Made in France.* Pour un artefact ne désignant pas un objet concret, la traduction de *faire* est variable : elle peut être *make,* suivi ou non d'une particule :

(7') *John (drew up + made up) a list of terms*
(8') *John (drew + made) a map of Mali*

5. Une personne lancée dans une vaste entreprise de déchirage de livres peut éventuellement dire à un de ses collègues *Ça y est, j'ai fait un livre sur ce sujet* pour indiquer qu'elle vient de le déchirer. Mais cette interprétation relève d'un contexte tellement marqué qu'elle peut être laissée de côté.

mais elle peut aussi être éventuellement *do :*

(5'a) *John (wrote + ?did) a book on this topic*
(6') *John (composed + wrote + produced + ?did) four symphonies*
(9') *John (painted + ?did) a portrait of Mary*

ou un autre verbe à usage multiple comme *have* ou *take* :

(10') *John (had + *did + *made) a piss*
(11') *John (took + *did + *made) a picture of Mary*

Il semble qu'il n'y ait pas de régularité totale quant à la traduction de *faire* dans ces cas-là : ainsi, on peut se demander pourquoi dit-on *make a soup* et *do a roast ?* pourquoi dit-on *(make + do) up a parcel ?*

Examinons comment obtenir automatiquement la traduction de la construction (A) (dans un système à transfert) en supposant que l'on ait identifié que N_0 est l'auteur de N_1 (ce qui n'est pas toujours évident comme nous le verrons dans les sections suivantes). Nous allons proposer une solution qui s'inspire de la « structure de qualia » proposée par Pustejovsky pour un lexique génératif (Pustejovsky 1991). L'une des hypothèses mise en avant par Pustejovsky est que de nombreuses informations lexicales doivent être codées sous les noms désignant des artefacts (et pas exclusivement sous les verbes) et que ces informations doivent interagir dynamiquement avec celles codées sous les verbes afin de calculer le sens d'une phrase. Cette hypothèse converge avec la solution que nous avons proposée pour les constructions à verbe support (Danlos 1992 ; Danlos et Samvelian 1992) dans la mesure où une grande importance est accordée aux informations lexicales codées pour les noms, ces informations guidant les processus d'analyse et/ou de traduction automatique.

Dans la structure de qualia élaborée par Pustejovsky, sont enregistrées entre autres les informations suivantes pour un nom désignant un artefact[6] :

- le rôle tellique de l'artefact qui est défini par un verbe, par exemple : *manger (eat)* pour *gâteau (cake), lire (read)* pour *livre (book),*
- le rôle agentif qui est défini par le(s) verbe(s) spécifique(s) décrivant la création de l'objet, par exemple : *préparer (bake)* pour *gâteau (cake), écrire (write)* pour *livre (book).*

Pustejovsky se sert de ces informations, entre autres, pour calculer la différence de sens entre *bake a (cake + potato)* – respectivement création d'un objet et changement d'état d'un objet – sans postuler deux entrées de *bake* mais en utilisant le fait que ce verbe est enregistré sous le rôle agentif de *cake* et non sous celui de *potato.* Nous proposons d'ajouter dans le rôle agentif d'un nom français ou anglais désignant un artefact le pro-verbe pouvant se substituer au(x) verbe(s) spécifique(s) de création, si nécessaire. Ainsi le nom *gâteau (cake)* peut être codé avec *préparer (bake)* comme verbe spécifique de création et *faire (make)* comme pro-verbe de création ; de même, *livre (book)* peut être codé avec *écrire (write)* comme verbe de création et *faire (do)*

6. Ces informations lexicales sont aussi *grosso modo* codées dans le dictionnaire de Mel'čuk (1984) où elles sont indexées sous des rubriques appelées « fonctions lexicales ».

comme pro-verbe de création. Rappelons que nous ne nous aventurons pas à poser que tous les noms désignant un artefact acceptent *faire* comme pro-verbe de création (c'est-à-dire, entrent dans la construction (A) avec l'interprétation où N_0 être l'auteur de N_1). S'il s'avérait qu'une sous-classe bien définie de ces noms possède cette propriété, une règle générale pourrait remplacer un codage systématique sous chaque nom de cette classe, mais ceci relève d'une question d'ergonomie de codage et n'affecte en rien les principes de construction de lexique et de processus de traduction que nous avançons ici.

Examinons l'utilisation de ces informations lexicales dans un système de traduction automatique à transfert. Pour les phrases construites avec un verbe spécifique de création, la traduction est apparemment compositionnelle :

Tra(Jean a écrit un livre sur ce sujet)
= Tra(Jean) Tra(a écrit) Tra(un livre) Tra(sur ce sujet)
= John wrote a book on this topic

ce qui peut laisser à penser que l'information que *écrire (write)* est le verbe spécial de création de *livre (book)* est superflue. Cependant, considérons la paire suivante :

(14) Jean a établi la facture
(14') *John wrote (out) the bill*

Elle ne relève pas d'une traduction compositionnelle dans la mesure où *établir* et *write (out)* ne sont pas mis en correspondance hors de ce contexte particulier, c'est-à-dire avec comme objet *facture* ou *bill*. Pour obtenir la traduction de (14) en (14') – ou de (14') en (14) – nous proposons la méthode suivante tout à fait similaire à celle que nous avons élaborée pour les constructions à verbe support : dans la phase d'analyse, identifier l'emploi de *facture*[7] et identifier que *établir* est codé comme verbe de création de *facture* ; dans la phase de transfert bilingue, chercher la traduction de *facture* dans l'emploi concerné, c'est-à-dire, *bill*, et chercher le verbe spécial de création de *bill*, c'est-à-dire, *write (out)* ; dans la phase de génération, construire une phrase anglaise avec *write (out)*. Cette méthode évite une règle lexicale comme *établir (- facture)* ---> *write (out)* qui est *ad hoc*.

Tournons-nous maintenant vers les phrases construites avec le pro-verbe de création *faire*. La paire suivante :

(15) Jean a fait la facture
(15') *John made (out) the bill*

peut être obtenue d'une façon tout à fait similaire à celle décrite pour la paire (14)-(14') : il suffit de travailler sur le pro-verbe de création de *facture* et de *bill* au lieu de travailler sur leur verbe spécifique de création. Notons que l'on peut aussi travailler sur le pro-verbe de création dans la langue source et sur le verbe spécifique de création dans la langue cible afin d'obtenir une paire comme :

7. Le nom *facture* a au moins un autre emploi illustré dans l'exemple *Cet artiste a une facture personnelle* (—> *This artist has a personal technique*).

(15) Jean a fait la facture
(14') *John wrote (out) the bill*

En conclusion, bien que la traduction du pro-verbe de création *faire* dépende de l'objet direct qu'il introduit, la méthode de transfert que nous proposons permet d'éviter des règles contextuelles coûteuses telles que *faire (- facture) —> make (out).* Cette méthode est de plus facilement implémentable dans la mesure où elle ressemble à celle qui a déjà été implémentée pour les constructions à verbe support. La difficulté réside en fait dans le module d'analyse (et de génération), plus précisément dans les questions de structure argumentale : par exemple, quelle structure argumentale doit-on donner aux verbes *écrire, imprimer, publier* d'une part, au pro-verbe *faire* d'autre part, et enfin au nom *livre* pour que les propriétés des phrases (5), (5a-c) et de toute phrase comportant le nom *livre* soient prises en compte, et ce de manière cohérente ? Une ébauche de solution écrite en HPSG *(Head-Driven Phrase Structure Grammar,* Pollard et Sag 1987 ; 1993) est à l'étude (Danlos et Samvelian, à paraître ; Samvelian, à paraître) en vue d'une implémentation en ALEP *(Advanced Linguistics Engineering Platform)* dans le prolongement du travail de (Namer et Schmidt 1993) sur les constructions à verbe support.

N_0 est l'auteur de N_1 ou N_0 pratique N_1

Un bon nombre de noms français désignent soit un objet concret soit une activité, et se construisent avec *faire* dans les deux interprétations :

Jean a (fabriqué + fait) des skis
Jean a fait du ski

Jean a (tricoté + fait) un tricot
Jean a fait du tricot

Lorsque *faire* introduit un nom d'activité, que celui-ci désigne aussi un artefact comme *ski* ou non comme *natation*, le déterminant introduisant N_1, qui est obligatoirement au singulier, est *de le/la* et éventuellement *un/une* accompagné d'un modifieur (Pivaut 1989) :

Jean a fait (du ski + ?un ski très sportif)
*Jean a fait (un + le + ?ce + des + les + ces) ski(s)

Jean a fait (de la natation + ?une natation très sportive)
*Jean a fait (une + la + ?cette + des + les + ces) natation(s)

Par contre, lorque *faire* introduit un artefact avec l'interprétation N_0 est l'auteur de N_1, le déterminant introduisant N_1 n'est contraint que par les propriétés comptable/massique de ce nom :

Jean a fait (une + la + cette + des + les + ces) robe(s)
*Jean a fait de la robe

Ces différences de déterminants permettent de désambiguïser la construction (A) lorsque N_1 désigne à la fois un artefact et un nom d'activité ; seul le cas où N_1 est introduit par *un/une* avec un modifieur reste ambigu :

Jean a fait un ski très sportif (= Jean a (fabriqué + pratiqué) un ski très sportif)

Signalons que ces données linguistiques impliquent qu'un module d'analyse du français doit avoir accès au déterminant de l'objet direct pour désambiguïser certaines phrases.

En anglais, il existe généralement deux mots distincts pour désigner un artefact et l'activité mettant en jeu cet artefact[8] :

(fabriquer + faire) des skis --> *make skis*
faire du ski --> *go skiing, do some skiing*

(tricoter + faire) un tricot --> *(knit + make) a sweater*
faire du tricot --> *do some knitting*

(dessiner + faire) un tatouage --> *(? + do) a tattoo*
faire du tatouage --> d*o some tattooing*

(fabriquer + faire) une voiture --> *(build + make) a car*
faire de la voiture --> *go driving, do some driving*

En traduction automatique, il semble nécessaire de créer deux entrées distinctes pour les noms français désignant à la fois un artefact et une activité, et d'associer à chacune de ces entrées sa traduction anglaise, par exemple : *ski (+ artefact)* --> *ski* et *ski (+ activité)* --> *skiing*. De même, il semble nécessaire de créer deux entrées de *faire* selon qu'il introduit un artefact ou un nom d'activité : la traduction de l'entrée de *faire* introduisant un artefact (avec le sens N_0 est l'auteur de N_1) se calcule comme décrit dans la première section, la traduction de *faire* introduisant un nom d'activité peut être *do* ou éventuellemnt *go* pour des noms désignant une activité sportive. La principale difficulté réside donc dans la désambiguïsation de la phrase française. L'affaire n'est cependant pas si simple. Considérons, par exemple, les noms désignant un instrument de musique. Ils désignent, en français comme en anglais, soit un artefact soit l'activité artistique consistant à pratiquer cet instrument :

(fabriquer + faire) une trompette ---> *(construct + make) a trumpet*
(jouer + faire) de la trompette --> *play the trumpet*

Pour être cohérent avec ce que nous avons présenté jusqu'ici, il faudrait créer, en français et en anglais, deux entrées pour chaque instrument de musique, l'une correspondant à l'artefact, l'autre correspondant à l'activité. Néanmoins, cette solution n'est pas forcément satisfaisante. Une solution concurrente consiste à ne considérer qu'une seule entrée pour un instrument de musique en enregistrant dans sa structure de qualia les verbes *jouer (play)* et *faire* pour le rôle tellique, et *fabriquer (build)* et

8. Cette situation s'observe aussi en français, par exemple avec le couple *patin / patinage*.

faire (make) pour le rôle agentif. Cette solution évite, semble-t-il à juste titre, un dédoublement d'entrées pour les instruments de musique, mais elle n'est pas cohérente avec celle proposée pour des noms comme *ski, tricot, tatouage* ou *voiture* pour lesquels le dédoublement d'entrées semble inévitable. Nous ne trancherons pas entre ces deux solutions, car la situation est en fait encore plus compliquée, comme le montrent les exemples suivants :

(fabriquer + faire) une poterie --> *make (a piece of + *a) pottery*
faire de la poterie ---> *do some pottery*

Combien d'entrées faut-il créer pour le nom *pottery* ? La réponse à cette question demande d'aborder des problèmes délicats de détermination, ce qui sort largement du cadre de cet article. Nous nous contenterons d'émettre l'avis suivant : quelle que soit la ou les solution(s) adoptée(s) pour traiter des noms comme *ski (ski, skiing), trompette (trumpet)* et *poterie (pottery)* dans leurs constructions avec *faire* (un verbe anglais), on peut s'attendre à ce que cette ou ces solution(s) ne soi(en)t pas satisfaisante(s) sur tous les plans, ce qui implique, entre autres, de renoncer à une certaine « élégance » dans la construction du lexique ou l'élaboration d'un formalisme grammatical, même basé sur une grammaire lexicalisée.

Interprétations diverses de la construction (A)

Il existe un certain nombre de phrases de structure (A) qui peuvent être qualifiées de figées et qui reçoivent des traductions anglaises diverses :

(1) faire le mur --> *leap over the wall*
faire les carreaux/la vaisselle --> *do the window panes/the dishes*
faire un lit --> *make up a bed*
faire ses valises --> *pack one's bags*[9]
faire les magasins --> *do the stores*
faire des mots croisés --> *do crosswords*
faire un problème de maths --> *do a Maths problem*

Ces expressions peuvent être qualifiées de figées dans la mesure où la relation entre faire et N_1 est opaque et où le déterminant introduisant N_1 peut être contraint :

faire (le + *un + *son +*ce +*les + *des +*ses +*ces) mur(s)

bien que ce ne soit pas toujours le cas :

faire (le + un + son + ce + les + des + ?*ses + ces) lit(s)

Les phrases construites autour de ces expressions sont *a priori* toutes ambiguës : elles peuvent recevoir l'interprétation figée et l'interprétation N_0 est l'auteur de N_1. Néanmoins, il semble que cette ambiguïté soit souvent plus théorique que réelle : dans les textes, on évitera généralement d'employer *faire* lorsque l'on veut l'interprétation

9. L'expression *faire ses valises* a deux interprétations : remplir ses valises et partir. Dans ces deux interprétations, elle se traduit par *pack one's bags* en anglais.

N_0 est l'auteur de N_1 en ayant recours à un verbe de création spécifique, comme *construire le mur* ou *concevoir des mots croisés*[10]. En d'autres termes, on peut ne retenir que l'interprétation figée de ces expressions et en traduction automatique les traiter comme des cas figés. Cette solution implique que les noms d'artefacts concernés ne soient pas marqués comme acceptant le pro-verbe de création *faire*.

En dehors de ces cas figés qui sont listables, le verbe *faire* dans la construction (A) peut être employé, surtout à l'oral, à la place de verbes variés :

(2) Jean a (étudié + fait) ce roman --> *John (studied + did) this novel*
Jean a (visité + fait) cette cathédrale ---> *John (visited + did) this cathedral*

Il semble qu'il se traduise alors sytématiquement par *do*. À niveau de langue égal, c'est-à-dire, dans un style relâché, les phrases (2) avec *faire* sont toutes ambiguës : elles ont l'interprétation indiquée et l'interprétation N_0 est l'auteur de N_1. Contrairement aux phrases figées de (1), il n'y a pas d'interprétation préférentielle à niveau de langue égal : la seule façon de désambiguïser consiste à faire appel à des connaissances contextuelles et/ou extralinguistiques (par exemple : est-ce que Jean est un romancier ou un étudiant ?). Une fois la désambiguïsation de la phrase en *faire* faite, sa traduction ne pose pas de problème particulier. Dans un style soutenu, on peut considérer que les phrases (2) avec *faire* n'apparaissent pas avec le sens indiqué.

Il existe encore d'autres ambiguïtés de phrases de structure (A), mais nous n'en citerons qu'une dernière : pour tout nom désignant un artefact qui se vend, *faire* peut être employé à la place de *vendre :*

(3) Jean (vend + fait) cette marque de gants --> *John (sells + does) this make of gloves*

Là aussi, seul le contexte permet de désambiguïser une phrase comme (3) avec *faire*, qui, une fois désambiguïsée, ne pose pas de problème de traduction particulier.

Conclusion

On retiendra que la structure (A) *N_0 (+ humain) faire N_1 (+ artefact)* peut recevoir des interprétations variées et des traductions anglaises tout aussi variées. Une des principales difficultés pour traiter les phrases de structure (A) consiste donc à lever les ambiguïtés, ce qui est un problème classique[11]. Pour les phrases de structure (A) qui ont l'interprétation N_0 être l'auteur de N_1, nous avons proposé une solution qui s'inspire à la fois de la structure de qualia élaborée dans Pustejovsky (1991) et du traitement des constructions à verbe support élaboré dans Danlos (1992) et dans Danlos et Samvelian (1992). Cette solution demande de coder beaucoup d'informations lexi-

10. Les expressions *faire un bon dîner/repas* semblent cependant s'employer tant dans le sens création que dans le sens figé de *manger* orienté vers la fonction tellique de ces noms. Elles se traduisent dans les deux sens par *make a good dinner/meal*.
11. Vu les difficultés évoquées, nous (suggérons + avançons + faisons) la recommandation suivante : éviter, si possible, d'employer *faire* dans des textes qui doivent être analysés ou traduits automatiquement.

cales sur les noms désignant un artefact, mais ceci est un point de passage obligé si on veut obtenir la traduction des phrases (A) ainsi que celle des phrases où un verbe spécifique de création remplace *faire* (par exemple : *Jean a établi la facture--> John wrote (out) the bill).* Pour les noms désignant à la fois un artefact et une activité (par exemple : *ski (ski, skiing), trompette (trumpet), poterie (pottery)),* il est clair qu'il faut aussi coder beaucoup d'informations lexicales, mais l'organisation de ces données lexicales ne va pas de soi : il est possible que l'on soit amené à renoncer à la construction d'un lexique « élégant » pour ces noms.

Références

BOONS, J. P., GUILLET, A. et Ch. LECLÈRE (1976) : *Structures des phrases simples en français: classes de constructions transitives,* Rapport de Recherches du LADL n° 6.

DANLOS, L. (1992) : « Support Verb Constructions: Linguistic Properties, Representation, Translation », *Journal of French Language Studies,* vol 2, n° 1, Cambridge, Cambridge University Press.

DANLOS, L. et P. SAMVELIAN (1992) : « Translation of the Predicative Element of a Sentence: Category Switching, Aspect and Diathesis », *Proceedings of the Fourth International Conference on the Theoretical and Methodological Issues in Machine Translation of Natural Languages* (TMI-92), Montréal, CCRIT.

DANLOS, L. et P. SAMVELIAN (à paraître) : « Les verbes supports: pour un outil indispensable et/ou contre une notion imprécise ? »

GIRY-SCHNEIDER, J. (1978) : *Les nominalisations en français: l'opérateur « faire » dans le lexique,* Genève, Droz.

GIRY-SCHNEIDER, J. (1987) : *Les prédicats nominaux en français: les phrases simples à verbe support,* Genève, Droz.

MEL'ČUK, I. (1984) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches Lexico-Sémantiques I,* Montréal, Presses de l'Université de Montréal, 172 p.

PIVAUT, L. (1989) : *Verbes supports et vocabulaire technique,* Thèse de Doctorat, Université Paris VII.

POLLARD, C. et I. SAG (1987) : *Information-Based Syntax and Semantics, Vol 1: Fundamentals,* CSLI Lecture Notes n° 13, Standford.

POLLARD, C. et I. SAG (1993) : *Head-Driven Phrase Structure Grammar,* University of Chicago Press.

PUSTEJOVSKY, J. (1991) : « The Generative Lexicon », *Computational Linguistics,* vol 17, n° 4, MIT Press.

SAMVELIAN, P. (à paraître) : *Les nominalisations en français: structure argumentale et réalisation actancielle,* Thèse de Doctorat, Paris, Université de Paris VII.

NAMER, F. et P. SCHMIDT (1993) : « Une grammaire du français dans un formalisme à structure de traits typés », Actes *Informatique et Langage Naturel (ILM),* Nantes.

8

La TAO comme technologie scientifique : le cas de la TA fondée sur le dialogue

Christian BOITET

GETA, IMAG, Université Joseph-Fourier et CNRS, Grenoble, France

• ***Abstract*** •

Opinions on Machine Translation (MT) are often extreme. Some consider it only as a testbed for the scientific experimentation of their favorite theories or formalisms, while others see it as a purely technological and utilitarian enterprise. In most cases, each defends a particular "paradigm", as one would defend an ideology.

For example, it is asserted that progress could only come by building systems equipped with an ontology, which "understand" explicitely, while this is rarely possible, and, if it is possible, rarely necessary. Or it is asserted that it is necessary to go through an interlingua in order to build multilingual systems, while this already old idea is particularly difficult to implement, and other approaches are equally possible and considerably less expensive, except in the rare situations which are strongly multilingual (at least 8 languages, with balanced throughputs for all language pairs).

Our thesis is that MT is neither a science nor a collection of recipes, but rather a scientific technology, that is, a set of methods which progress as much through episodic integration of theoretical ideas as through incremental improvements in practical know-how. On the other hand, the famous "paradigms" cannot be ranked on a unique scale of values, but should be compared with reference to each type of translational situation. Finally, as far as the linguistic architecture of MT systems is concerned, there remain many intermediate avenues to explore.

This thesis can be illustrated with the new paradigm of "Dialogue-Based MT", or personal MT for monolingual authors, which we propose for the case of situations where other approaches, such as Language-Based MT and Knowledge-Based MT, are inadequate. Although the linguistic knowledge base of the system remains crucial, and should even be as covering as possible, and although extralinguistic knowledge may possibly be used if it is available, emphasis is on an indirect preedition, realized through a normalization and clarification dialogue with the author, making it possible to produce high quality translations without postediting.

In the last part of the paper, we present the specification and the current state of our mockup, LIDIA-1, first step towards building a DBMT system for a particular situation, that of monolingual authors writing hypertext technical documentation to be distributed in several languages. We also mention some interesting and new problems which have appeared during this work, problems for which we should find at least partial solutions before we could go further and build an operational prototype.

Introduction

Les avis sur la traduction automatique (TA) sont souvent extrêmes. Les uns ne la conçoivent que comme l'expérimentation scientifique de leurs théories ou de leurs formalismes favoris, tandis que les autres y voient une entreprise purement technologique et utilitaire. Le plus souvent, chacun défend un *paradigme* particulier, comme on défendrait une idéologie. Par exemple, on soutient que le progrès ne pourrait venir qu'en construisant des systèmes munis d'ontologie, qui *comprennent* explicitement, alors que c'est rarement possible, et, quand ça l'est, rarement nécessaire. Ou encore, on affirme qu'il est nécessaire de passer par un interlingua pour construire des systèmes multilingues, alors que cette idée déjà ancienne est particulièrement difficile à mettre en œuvre, et que d'autres approches sont également possibles et nettement moins coûteuses, sauf dans les rares situations très fortement multilingues (au moins 8 langues, avec flux équilibrés pour toutes les paires (Boitet 1988a et b ; Boitet 1990a).

Notre thèse est que la TAO n'est ni une science, ni une collection de recettes, mais plutôt une technologie scientifique, c'est-à-dire un ensemble de méthodes qui progresse autant par l'intégration épisodique d'idées théoriques que par l'amélioration incrémentale des savoir-faire. D'autre part, les fameux *paradigmes* ne peuvent être placés sur une seule échelle de valeurs, mais sont à comparer dans chaque type de situation traductionnelle. Enfin, chaque paradigme correspond à un assez grand nombre de choix fondamentaux sur l'architecture linguistique d'un système de TAO, et il reste de nombreuses voies intermédiaires à explorer.

Cette thèse peut être illustrée par le paradigme de la *TA fondée sur le dialogue* (TAFD), ou TAO personnelle pour auteur monolingue, que nous proposons pour le cas de situations traductionnelles où d'autres approches, telles que la TA fondée sur la langue (TAFL), la TA fondée sur la connaissance (TAFC), et les aides informatisées au traducteur (THAM), sont inadéquates. Dans cette approche, bien que la base de connaissances linguistiques du système reste cruciale, et doive même être la plus couvrante possible, et que des connaissances extralinguistiques puissent éventuellement être utilisées si elles sont disponibles, l'accent est mis sur une préédition indirecte, effectuée grâce à un dialogue de normalisation et de clarification avec l'auteur visant à produire des traductions de haute qualité sans révision.

Nous étudions diverses questions relatives à cette approche grâce à une maquette, LIDIA-1, première étape vers la construction d'un système de TAFD pour une situation particulière, celle d'auteurs monolingues rédigeant en hypertexte de la documentation technique devant être diffusée en plusieurs langues. Nous terminons en évoquant quelques problèmes intéressants et nouveaux apparus durant ce travail, et auxquels il faudrait au moins trouver des solutions partielles pour passer ultérieurement à un prototype.

TAO et IA : un système de TAO ne peut et ne doit souvent pas comprendre explicitement

Un système de traduction automatique doit-il et peut-il comprendre ? Si cette question est encore vivement controversée, 45 ans après les débuts de la TA, c'est parce que l'analyse se place à un niveau trop *angélique*, trop anthropomorphique, et trop éloigné des réalités techniques.

On pose souvent comme acquis qu'un traducteur humain peut et doit comprendre pour bien traduire, et on prend comme étalons la qualité de ses traductions, supposée parfaite, et le degré de sa compréhension, supposé complet. Mais il faut s'interroger sur ces prémisses, car on peut aussi bien soutenir qu'on ne peut traduire sans comprendre que démontrer qu'à vouloir trop comprendre, on ne peut plus traduire. On doit alors admettre que, selon la qualité de traduction recherchée, le traducteur ou l'interprète *doit* plus ou moins comprendre, et qu'en fonction de sa formation antérieure, de l'information accessible et des délais imposés, il *peut* plus ou moins comprendre.

Poser cette question suppose aussi que les systèmes de TA soient faits pour *remplacer l'homme,* c'est-à-dire pour fournir des traductions *équivalentes à celles que l'on demande aux humains.* Mais, en entrant un peu dans le détail des types d'automatisation de la traduction, on s'aperçoit que cette supposition est bien souvent erronée. Traduction humaine et TAO ne visent souvent pas la même chose, et la *compréhension* d'un système ne peut se définir ni se juger comme celle d'un humain. Pour cela, il convient d'étudier les différentes architectures envisageables pour les systèmes de TAO, et de déterminer le ou les *degrés* de *compréhension* qu'elles autorisent (compréhension explicite, apparente directe ou indirecte, implicite).

En examinant ce qui existe ou est réalisable actuellement, et ce qui reste du domaine des perspectives à plus long terme, on arrive alors à la conclusion suivante : dans certains cas, les systèmes de TAO peuvent comprendre au sens fort (explicitement), mais il n'est pas sûr qu'ils le doivent jamais, car des systèmes utilisant les autres degrés de compréhension peuvent souvent produire des traductions de qualité équivalente, pour un coût bien moindre.

Compréhension et traduction humaine

Très souvent, on entend des arguments du type : « un traducteur ne peut bien traduire que ce qu'il comprend » ou « les interprètes (simultanés) traduisent bien sans avoir le temps de comprendre » . Qu'entend-on donc par *bien traduire* ? (Et ce n'est déjà pas le même *bien* dans les deux assertions précédentes !)

Bien traduire, c'est en premier lieu transmettre le contenu objectif d'un message (ce qui est dit d'une réalité externe, concrète ou abstraite – contenu propositionnel, et comment cela est dit – modalité, type de discours, situation de communication...). En second lieu, c'est aussi rendre ses aspects plus subjectifs (style, tonalité affective, environnement culturel, aspects esthétiques ou rhétoriques, intentions cachées...).

On emploie le terme de *traduction* aussi bien pour la poésie que pour les ro-

mans, les rapports et manuels techniques, et les nomenclatures de pièces détachées, alors qu'il conviendrait, au moins, de distinguer entre :

- la *re-création,* par exemple la traduction d'Edgar Poe par Baudelaire, qui vise avant tout à transmettre l'aspect subjectif, fût-ce au prix d'une légère transformation du contenu ;
- la *localisation,* largement pratiquée pour les manuels de micro-ordinateurs, qui vise à adapter un contenu à un environnement culturel particulier ;
- la *traduction-diffusion* (Bourbeau 1990), en particulier la traduction de documentations techniques dont le contenu doit être strictement rendu, sans ajout ni omission, même si le style *sent la traduction ;*
- la *traduction rapide* enfin, dans laquelle nous rangerons la *traduction-dépistage* de textes écrits et l'interprétation simultanée.

La même traduction pourra donc être jugée *bonne* en traduction rapide, et détestable en re-création. À l'évidence, le traducteur humain qui effectue la localisation d'un manuel informatique comprend plus profondément qu'un interprète qui traduit des interventions techniques sur la politique agricole commune.

Compréhension et traduction : le mieux est parfois l'ennemi du bien !

Il semble même, à l'expérience, qu'à chaque type de traduction corresponde un *degré de compréhension* (humaine) optimal, et en particulier que, comme en bien d'autres domaines, le mieux puisse ici être l'ennemi du bien.

Lors du colloque COLING-73, à Pise, on put en voir une démonstration éclatante. Un Soviétique, le P[r] Andreev, devait présenter une communication en russe, et le D[r] Andreewski, d'origine russe et parfaitement à l'aise dans les deux langues, avait accepté d'effectuer une interprétation *différée* (traduction par morceaux de quelques minutes de parole). Le malheur voulut qu'il connût aussi parfaitement le sujet de l'exposé : avant de traduire le premier fragment, il interrogea l'orateur pour être sûr d'avoir bien compris un point de détail, et l'exposé se transforma rapidement en un dialogue en russe, d'où sortaient de temps à autre quelques phrases en français, hors contexte et donc sybillines... La traduction avait été tuée par la compréhension, alors qu'un interprète professionnel, totalement incompétent en linguistique quantitative et en statistique, aurait produit une traduction tout à fait acceptable après une étude convenable de la terminologie utilisée.

Un autre danger qui guette le traducteur est celui de chercher l'élégance en croyant comprendre, alors qu'il n'a pas la compétence nécessaire, et de commettre alors beaucoup plus de contresens que s'il se limitait à une traduction plus littérale. Ainsi, en 1977, nous avions rédigé une communication en français au colloque *Franchir la barrière linguistique* organisé par la CEE à Luxembourg. Comme il fallait la traduire en anglais et en allemand, deux chercheurs d'origine anglaise et allemande avaient envoyé toute la terminologie nécessaire dans les trois langues. Les deux premiers jets de traduction furent absolument incompréhensibles, les traducteurs n'ayant pu résister au désir de manifester qu'ils comprenaient ce qu'ils traduisaient, et d'améliorer les traductions proposées par nos collègues. Il fallut à peu près tout refaire.

Enfin, il faut voir que la traduction humaine de haute qualité est souvent produite par plusieurs personnes. Typiquement, le traducteur qui effectue le premier jet a une compétence technique très superficielle, mais connaît très bien la terminologie et les deux langues. Le premier réviseur est un traducteur *senior,* spécialiste du type de document considéré, et à même d'assurer l'homogénéité terminologique et stylistique. Enfin, on fait parfois intervenir un second réviseur, spécialiste du contenu technique du document et éventuellement ignorant de la langue source, pour détecter les contresens sémantiques linguistiquement plausibles et d'éventuelles ambiguïtés accidentellement introduites en langue cible. Chacun de ces intervenants *comprend,* certes, mais de manière différente.

Quelles connaissances doit-on utiliser pour « bien traduire » ?

Il y a essentiellement trois grands types de connaissances qui interviennent dans l'activité de traduction. Elles concernent le texte (linguistique) et le contexte, à savoir, l'univers de référence (sémantique) et l'aspect communicatif (pragmatique).

La connaissance linguistique ne se limite pas seulement à la connaissance des deux langues en présence. Elle concerne aussi les particularités du type de documents à traduire (vocabulaire, grammaire). Ainsi, il faut une assez longue expérience pour bien traduire des bulletins météo.

En général, un traducteur humain comprend facilement le contexte pragmatique et communicatif d'un texte. Par contre, sa connaissance du domaine spécialisé concerné est souvent très superficielle, voire nulle. Pourtant, s'il est expérimenté, il arrive dans une large mesure à faire illusion, c'est-à-dire à traduire comme s'il avait compris, en se fondant uniquement sur son habitude du type du texte. Il s'agit donc de *compréhension apparente* (humaine). L'absence de compréhension réelle se manifeste alors dans des fautes particulières. Prenons un exemple en théorie des langages formels et des automates, en supposant l'original en russe (où il n'y a pas d'articles).

исходя из правой линеарной грамматики г1, строят ассоциированную систему уравнений, и выводят регулярное выражение л(г1).	*À partir de* ***la*** *grammaire linéaire droite G1, on construit* ***le*** *système d'équations associé, et on en déduit* ***une*** *expression régulière pour L (G1).*

On peut *deviner* qu'il s'agit probablement de l*a* grammaire et non d'*une* grammaire à cause de la référence à G1, si G1 apparaît auparavant. Par contre, pour ne pas traduire par ***un*** *système d'équations* ni ***l'****expression régulière,* il faut impérativement connaître la théorie en question. En fait, il n'est pas nécessaire de comprendre pourquoi il s'agit *du* système et d'*une* expression, il suffit de le savoir.

Qu'il s'agisse de compréhension explicite ou de compréhension apparente, il est donc possible de parler de *niveaux* de compréhension chez le traducteur humain, en fonction de son degré de connaissance du domaine et d'habitude des particularités des textes. Enfin, rien n'interdit à un traducteur – et c'est même fortement conseillé – de se renseigner auprès d'un collègue plus expérimenté, d'un spécialiste du domaine, ou de l'auteur. Qu'elle soit explicite ou apparente, sa compréhension est alors *indirecte.*

En pratique, des compromis sont nécessaires

Selon la qualité de traduction recherchée, le traducteur ou l'interprète *doit* plus ou moins comprendre. Pour une re-création ou une localisation, il faut une compréhension réelle, explicite, assez profonde. Par contre, pour une traduction-diffusion ou une traduction rapide, une compréhension apparente est suffisante. Notons toutefois que, dans le premier cas, la traduction doit être bien meilleure que dans le second, pour qu'un réviseur acccepte de réviser. Dans le second, une traduction purement littérale peut être utile à un utilisateur pressé d'accéder à une information peu fouillée (par exemple, y a-t-il eu une expérience ou un brevet sur tel sujet au Japon ?).

Il ne faut cependant pas oublier que la traduction est une activité soumise à bien des contraintes. Un traducteur est souvent obligé d'accepter des traductions dans des domaines qu'il connaît peu, sur des textes divisés en petits morceaux et donc difficilement compréhensibles, et avec des délais toujours trop courts. En fonction de sa formation antérieure, de l'information accessible et des délais imposés, il *peut* plus ou moins comprendre.

Niveaux de compréhension en TAO

Voilà donc un point de départ. Cependant, pour parler des systèmes informatiques, il faut être plus précis sur les notions de qualité et de compréhension. La qualité ne peut à notre avis s'évaluer intrinsèquement, mais doit l'être par rapport à l'usage qu'on veut faire des traductions fournies.

Pour déterminer si la compréhension d'un système est explicite ou apparente, il convient d'analyser sa structure, et non son comportement : contient-il, oui ou non, un composant qui modélise l'univers de référence et la situation de communication, une *ontologie autonome ?*

De quelle TAO parle-t-on ?

Vers 1949, les USA, puis l'URSS, ont lancé des programmes motivés par le besoin de renseignements. C'est la *TAO du veilleur.* Il s'agit de traduction totalement automatique, dont on attend des traductions *grossières*, produites rapidement, en grand volume et à bas coût. La qualité de ces traductions n'est pas essentielle, car elles servent à filtrer des documents, dont les plus intéressants peuvent, si nécessaire, être traduits ou communiqués à des spécialistes bilingues. Préédition et postédition doivent être absentes ou très limitées (exemple : séparer les phrases, les formules, les figures...)[1]. Ce besoin est toujours actuel. Cependant, il s'agit maintenant plus de veille scientifique, technique, économique et financière que de renseignement militaire[2].

1. Les systèmes Systran sont essentiellement de ce type (par exemple, le système russe-anglais installé depuis 20 ans à la Wright-Patterson Air Force Base traduit, d'après nos informations, environ 18 millions de mots par an, avec une qualité tout à fait satisfaisante pour l'usage visé).
2. À titre d'exemple, on peut citer l'accès en anglais à des bases de données japonaises depuis la Suède (Sigurdson et Greatex 1987), ou encore depuis l'Europe (service Japinfo utilisant des traductions « grossières » d'ATLAS-II à peine « arrangées »).

Une quinzaine d'années plus tard, on a commencé à travailler sur la *TAO du réviseur*, dans laquelle on produit des traductions *brutes*, destinées à être révisées. Dans cette optique, la machine doit remplacer le traducteur, promu réviseur[3]. Typiquement, en moyenne industrielle, une page technique de 250 mots est traduite en 1 heure et révisée en 20 minutes. Avec 4 personnes, on passerait donc de 3p/h à 12p/h, et on multiplierait donc la productivité par 4. Il s'agit d'une limite, le chiffre le plus vraisemblable étant plutôt de 8 p/h, en comptant une révision plus lourde, de 30 mn/p.

Que faire pour la plus grande partie des textes dont on veut obtenir de bonnes traductions ? La bureautique a commencé à apporter des solutions, sous forme d'outils de *TAO du traducteur.* Il s'agit de traduction humaine assistée par la machine, ou THAM, et pas de traduction automatique. C'est l'utilisateur qui traduit, en s'aidant de dictionnaires bilingues, de bases terminologiques, de thesaurus de *bitextes* (textes + traductions), etc., accessibles depuis un traitement de texte, le tout formant un *poste de travail pour le traducteur,* réalisé sur micro-ordinateur ou station de travail. Il s'agit d'outils comme Mercury/Termex™ sur PC, WinTool™ sur Macintosh, ou de systèmes complets (Weidner-Bravice, TSS de Alps). Le réviseur peut utiliser le même environnement, ou préférer travailler directement sous l'outil final de PAO (publication assistée par ordinateur).

Pour la majorité des besoins, et en particulier pour la traduction de manuels d'enseignement dans des pays où la langue nationale ne s'est que récemment affirmée comme support de l'enseignement secondaire et universitaire, la THAM est actuellement la seule voie réaliste. Il en est de même de toutes les traductions scientifiques et techniques de faibles volumes homogènes, voire de grands volumes homogènes trop mal rédigés (résumés avec des phrases de 15 lignes, par exemple) ou non disponibles sous forme magnétique cohérente et sans erreurs.

Enfin, on commence à voir apparaître des outils de *TAO de l'auteur,* destinés à des personnes ignorant la langue source. Cela correspond à des besoins croissants. Pour l'instant, il s'agit en fait de textes multilingues préenregistrés, personnalisables grâce à des parties variables. Par exemple, Ambassador[TM*], disponible en anglais-japonais, anglais-français, anglais-espagnol et français-japonais, offre environ 200 *formats* de lettres et formulaires, et 450 *moules* de texte (phrase ou paragraphe). Pour ce qui est de textes libres, il n'y a pas encore de produits commerciaux (bien que le système JETS d'IBM-Japon (Maruyama, Watanabe et Ogino 1990) soit prêt au moins depuis 1992). En tout état de cause, il ne pourra s'agir que de systèmes fondés sur le dialogue (TAFD), dont nous reparlerons plus loin.

Traduction humaine et TAO ne visent souvent pas la même chose

La question de savoir si un système de TAO doit ou peut comprendre ne s'applique évidemment pas aux systèmes d'aide au traducteur humain, mais seulement aux

3. La plupart des systèmes commerciaux récents visent ce créneau. On peut citer des systèmes japonais (AS-Transac de Toshiba, ATLAS-II de Fujitsu, PIVOT de NEC, HICAT de Hitachi, DUET de Sharp, SHALT d'IBM-Japon, PENSÉE de OKI, Majestic de l'Université de Kyoto et du JICST,...), ou américains (LOGOS, METAL), ou français (Ariane/aéro/F-E de SITE-B'VITAL, fondé sur les outils informatiques et les méthodes linguistiques du GETA). Il est aussi possible d'adapter des systèmes de conception plus ancienne à cet usage, si on les spécialise à un langage fortement contrôlé, comme cela a été fait chez Xerox avec Systran pour la traduction de notices d'anglais en 4 langues.

* Disponible sur Macintosh et PC, produit par Language Engineering Corp., Belmont, MA 02178.

trois autres types de systèmes, qui présentent une automatisation totale ou partielle de la *fonction traduisante* (production d'un premier jet, traduction grossière ou brute).

Avec cette restriction, il est donc clair que ni la re-création ni la localisation ne sont abordées dans les systèmes de TAO actuels ou à l'étude : ici, l'homme n'est pas en concurrence avec la machine. De plus, si l'on classe les systèmes de TAO par type d'utilisateur (TAO du veilleur, du réviseur, du traducteur, du rédacteur), on s'aperçoit que, dans deux cas sur quatre (veilleur, traducteur), on ne cherche pas à produire une traduction comparable à la traduction humaine.

Ainsi, un système de TAO n'est pas toujours fait pour *traduire,* au sens *humain* du terme ! Là encore, la machine ne va pas remplacer l'homme, mais couvrir des besoins non couvrables par des humains, soit qu'on ne puisse trouver assez de traducteurs pour traduire dans le délai imparti, ou en temps réel (bases de données), soit que leur coût soit de toutes façons trop élevé, soit encore qu'ils refusent un travail trop répétitif et inintéressant (bulletins météo, par exemple).

Architectures des systèmes de TAO et degrés de compréhension

Un système de TAO, ou de TALN en général, peut comporter trois composants distincts, linguistique, sémantique et humain, le premier étant seul nécessaire. Comme les logiciens, nous dirons qu'il ne peut y avoir compréhension au sens fort, ou compréhension explicite, que si le texte est interprété dans le composant sémantique (*ontologie* dans KBMT-89 (Nirenburg *et al.* 1989) et KANT (Nyberg et Mitamura 1992)). Le *niveau de compréhension* dépend alors du détail et de la complétude de cette ontologie par rapport à l'univers de référence réellement associé au texte, et de l'exactitude de l'analyse.

La *fonction traduisante* des systèmes de TAO du veilleur et du réviseur ne comporte qu'un composant linguistique, dans lequel on représente la connaissance linguistique de base, et les régularités spécifiques au type de textes visé. Cependant, comme les langues possèdent une sémantique *intrinsèque* (Desclés 1987), on peut coder certaines connaissances sur le domaine dans la base lexico-grammaticale (traits et grammaires sémantiques), et on peut lever un certain nombre d'ambiguïtés sémantiques et pragmatiques au moyen d'heuristiques linguistiques. Même dans ce dernier cas, la compréhension n'est qu'apparente, ou *implicite*, c'est-à-dire que le système ne fait que permettre au lecteur (utilisateur final ou réviseur) de comprendre, il ne comprend pas lui-même.

Quel type de compréhension peut-on envisager pour la TAO de l'auteur ? Elle ne permet aucune révision, et, sauf si l'on conçoit un système pour un groupe de spécialistes, elle ne permet pas non plus le recours à une base de connaissances : la compréhension ne peut être qu'apparente. Enfin, elle ne permet pas de se restreindre à un langage *contrôlé* (langage artificiel, non ambigu, à « consonance naturelle »), ni même à un *sous-langage* déterminé. C'est le grand public qu'il faut viser. Par conséquent, il semble difficile de s'appuyer sur les régularités formelles des textes pour obtenir la quasi-perfection nécessaire.

Est-ce la quadrature du cercle ? Oui, si l'on veut tout à la fois : du texte libre, pas de préédition, pas de postédition, et une grande qualité. Non, si l'on accepte un langage guidé, et une interaction avec l'auteur, permettant d'arriver :

- à une *standardisation* du texte à traduire, sous ses aspects lexicaux, grammaticaux et stylistiques ;
- à une *clarification* de ce texte, permettant de réduire les ambiguïtés lexicales, grammaticales et sémantiques (exemple : *Jean a acheté un livre à Pierre*).

Les systèmes de TAO personnelle seront donc essentiellement fondés sur le dialogue, même si, dans certains cas, on peut imaginer de les lier à un système expert du domaine traité. Si l'on ne peut coder la connaissance nécessaire à une compréhension explicite parfaite, on peut utiliser le dialogue pour « aller la chercher dans la tête de l'auteur » , et aboutir à une compréhension apparente *indirecte,* c'est-à-dire à une représentation linguistique *profonde* du texte de qualité et de finesse au moins égales à celles que pourrait produire un processus de compréhension explicite (au moins égales, car l'auteur sera toujours plus compétent que tout système expert sur ce qu'il veut dire).

Réalités et perspectives

Comment ces considérations s'appliquent-elles en pratique ? On peut le voir en étudiant quelques systèmes, prototypes ou maquettes, par rapport aux situations concrètes dans lesquelles on les utilise ou les utiliserait.

Il n'est peut-être pas inutile de préciser ici ce que nous entendons actuellement *par systèmes, prototypes et maquettes* en TAO.

systèmes[4] : Il s'agit d'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.

prototypes[5] : Il s'agit d'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, c'est-à-dire sur un flux de textes nouveaux.

maquettes[6] : Il s'agit de programmes développés pour l'expérimentation réduite de techniques ou d'architectures nouvelles.

La traduction automatique quasi-parfaite sans compréhension explicite est possible dans des cadres restreints (TITUS, METEO)

Depuis près de 15 ans, le système METEO (Chandioux et Guérard 1981 ; Chandioux 1988), issu des travaux du groupe TAUM de l'Université de Montréal, traduit

4. Comme (Par ordre alphabétique) ARIANE (R-F au GETA, F-E à SITE-B'VITAL), AS-TRANSAC (E-J/J-E, Toshiba), ATLAS (E-J/J-E, Fujitsu), DUET (E-J/J-E, Sharp), ENGSPAN (E-S, PAHO), HICAT (E-J/J-E, Toshiba), JETS (J-E, IBM-Japon), KANT (E-F.S à Caterpillar), LOGOS (D-E.F...), METEO (E-F), METAL (D-E..., Siemens), MU-MAJESTIC (E-J/J-E, JICST), PENSEE (E-J/J-E, OKI), PIVOT (E-J/J-E, NEC), SHALT-J (E-J, IBM Japon), SPANAM (S-E, PAHO), SUSY (R.F.R-D, IAI), SYSTRAN (CEE, Xerox...).

5. Comme ARIANE (pour le F-M à l'USM), ATLAS (pour diverses langues), ETAP (R-E/E-R à l'IPPI, Moscou), EUROTRA (CEE), ITS (BYU, Provo), JEMAH (USM, Penang), LMT (IBM-USA), METAL (pour le F-D à Liège, ou les autres paires en développement), SYGMART (E-S au CELTA, Nancy), ULTRA (UNM, Las Cruces), etc.

6. Comme DIALOG (Kitano, CMU), ELU (Bouillon et Estival, Genève), KBMT-89 (CMT, CMU), LIDIA-1, etc.

les bulletins météo d'anglais en français. Actuellement, le volume est d'environ 40 000 mots par jour, et le taux de correction par les réviseurs humains est maintenu inférieur à 3 % (en adaptant périodiquement le linguiciel à la dérive lexicale et grammaticale inévitable). On entend par là qu'il y a moins de 3 manipulations pour 100 mots, un remplacement comptant pour 2 (suppression suivie d'insertion). Six ou sept postes de traducteurs sont épargnés. Autrefois, les traducteurs fuyaient ce *purgatoire.* Maintenant, ceux qui restent n'ont plus que le travail intéressant de révision, et restent volontiers plus longtemps.

À l'Institut Textile de France, on rédige à l'aide du système TITUS (Ducrot 1982 et 1988) des résumés d'articles techniques dans un langage très fortement contrôlé, et on fabrique sur demande les résumés dans plusieurs langues à partir de la forme interne, seule conservée. La situation est différente de METEO, puisque l'auteur est forcé de rentrer dans un moule prédéfini, dans lequel toutes les ambiguïtés sont soigneusement éliminées, alors que les rédacteurs météo suivent plus ou moins des règles de rédaction : il s'agit d'un sous-langage observé et non d'une *pseudo-langue* construite, ce qui entraîne d'ailleurs la présence de nombreuses ambiguïtés.

Les grammaires et dictionnaires de METEO codent dans une certaine mesure la sémantique de la météo. Par exemple, on utilise des groupes du type *phénomène météo,* ou *évolution de situation* plutôt que groupe nominal, groupe verbal, etc.

Ces deux techniques sont excellentes dans le cas de domaines restreints. Cependant, elles ne sont ni extensibles ni portables. De plus, on trouve en pratique très peu de situations où l'on peut définir un langage contrôlé et imposer son usage, ou bien observer un sous-langage portant sur un domaine très particulier et écrire des grammaires et des dictionnaires *sémantiques* permettant la compréhension apparente. D'ailleurs, ces deux systèmes restent des cas isolés, malgré les efforts déployés depuis plus de quinze ans pour trouver des opportunités analogues.

L'approche par « sous-optimisation » permet d'obtenir une qualité suffisante pour la révision, en compréhension apparente

Les sytèmes actuels de *TAO du réviseur* ont une couverture trop large pour qu'on puisse, même à long terme, imaginer de les munir d'une ontologie. Ils ne peuvent donc comprendre au sens fort. Cependant, on peut obtenir d'excellents résultats en spécialisant leurs dictionnaires, leurs grammaires et leurs heuristiques de désambiguïsation à des sous-langages convenables.

Laurent Bourbeau et John Lehrberger (Lehrberger et Bourbeau 1988, Bourbeau 1990) parlent à ce propos de *sous-optimisation.* C'est la même idée qu'en IA : puisqu'on n'arrive pas à résoudre le problème dans toute sa généralité, on construit des sortes de *systèmes experts* de la traduction d'un certain type de textes. Bien sûr, on peut procéder de la même façon en TAO du veilleur, puisque tout gain de qualité se traduit par une plus grande satisfaction de l'utilisateur final.

Pour montrer ce qu'on peut obtenir par sous-optimisation, et illustrer les idées précédentes de façon plus précise, voici quelques exemples tirés de traductions brutes produites par le système Ariane/aéro/F-E de B'VITAL.

Il s'agit d'un système dédié à des manuels de maintenance en aéronautique, et écrit à l'aide d'Ariane-G5, le générateur de systèmes de TAO construit au GETA, et d'outils annexes, en particulier la base lexicale BDTAO, construits par B'VITAL.

Après essai, s'assurer du fonctionnement correct de l'ensemble raccord.	*After test, check that the coupling assembly works correctly.*

On remarque ici le passage d'un groupe nominal prépositionnel *du fonctionnement correct* à un groupe verbal, *that... works correctly,* avec pour corollaire le passage de l'adjectif *correct* à l'adverbe *correctly.* Ces transformations ne sont pas effectuées au transfert. C'est la première étape de la génération syntaxique qui, à partir de la *g-structure* (structure génératrice), considérée comme sous-spécifiée relativement aux fonctions syntaxiques et aux classes syntagmatiques et morphosyntaxiques, recalcule ces niveaux en fonction de l'objectif initial (ici, construire une phrase verbale), à partir des niveaux plus profonds (relations logiques à l'intérieur du cadre prédicatif strict, relations sémantiques pour les compléments circonstanciels).

Grâce à la notion d'unité lexicale (famille dérivationnelle comme *réparer, réparateur, réparation, réparable,* ou *utile, utilité, utilement*), le générateur sait, sans avoir besoin de consulter un dictionnaire, quels lemmes contient la famille dérivationnelle considérée, ce qui conditionne les paraphrases possibles. Ici, *fonctionnement* a été ramené à l'UL *fonctionner-V,* traduite par *work-V,* qui porte la potentialité de dérivation vers un nom d'action. C'est donc simplement l'ordre de préférence des règles de choix des catégories syntagmatiques qui provoque la construction d'une subordonnée plutôt que d'un groupe nominal (*the correct working of the coupling assembly*).

Porter sur celle-ci la date de la dernière réception ou révision.	*Write on this one the date of the last reception or of service.*

Porter est ici un verbe support, et *porter une date* est traduit par *to write a date* et non par *to carry a date,* grâce à un test effectué en transfert lexical sur les traits syntaxiques et sémantiques de l'argument 1 de *porter* (l'objet logique). Mais le système ne *comprend* pas ce qu'est une date, ni ce qu'est l'écriture.

Effectuer la vidange générale et la purge du carburant (voir chapitre 12).	*Drain in a general manner and bleed fuel (see chapter 12).*

Effectuer la vidange est traduit par le verbe simple *to drain,* grâce à la notion d'unité lexicale, et à l'organisation du transfert lexical. *Vidange* est ramené à *vidanger-V,* et cette unité lexicale donne en traduction un arbre dans lequel on code la possibilité de la présence d'un verbe support du genre de *effectuer, faire,* etc., verbe dont la traduction sera effacée au cours du transfert structural.

Le bouchon a pour but d'assurer la protection d'un raccord auto-obturable lorsque celui-ci n'est pas utilisé au sol ou en vol.	*The trap is used for carrying out the self-sealing coupling protection when this one is not used at the ground or in flight.*

Avoir pour but est reconnu comme un prédicat composé, avoir-but-V(x0,x1), qui est traduit par use-V(x1,x0), avec conversion d'arguments, ce qui explique la génération d'un passif.

Enduire légèrement le joint neuf de liquide d'utilisation.	*Slightly coat the new joint with operating fluid.*

La traduction des prépositions est toujours délicate. Il faut savoir si elles introduisent des arguments ou des circonstants. Ici, *enduire-V* est un prédicat à 3 arguments (qn enduit qn/qc de qc), le troisième étant introduit par *de*. L'analyseur préfère compléter le cadre argumentaire, et l'introducteur de cet argument pour *coat-V* est *with*.

Ouvrir progressivement le robinet (3), appliquer une pression jusqu'à 1,5 bar jusqu'à l'allumage du voyant lumineux DS2 et l'extinction du voyant DS1.	*Gradually open tap (3), apply a pressure up to 1.5 bar until the light DS2 switching on ((ignition)) and the signal lamp DS1 extinction.*

La préposition *jusqu'à* introduit ici deux circonstants. Ce qu'on traduit en fait, c'est la relation sémantique (ici, RS = LOC avec SEM = TEMPS et SLOC = QUA), précisée par la préposition et par les traits sémantiques du gouverneur (*head*) du groupe, ici PROCESSUS pour *allumer-V*, et du prédicat (appliquer-V).

Ouvrir progressivement le robinet (3) jusqu'à obtenir une pression de 9 bars.	*Gradually open tap (3) until a pressure of 9 bars is obtained.*

Aucune transformation explicite n'est effectuée. La proposition infinitive est rendue par une subordonnée par le simple fonctionnement du générateur, expliqué plus haut. Comme l'argument 0 (sujet logique) n'est pas exprimé, on génère un passif. Il s'agit d'une préférence de style, et on pourrait aussi bien générer *until one obtains a pressure of 9 bars,* ou *until obtaining...*, comme dans l'exemple suivant.

Procéder à la dépose des panneaux.	*Remove the panels.*
IMPORTANT : avant de déposer ou de reposer le panneau central intrados de voilure, il est nécessaire de procéder à certaines modifications.	*IMPORTANT : before removing or reinstalling the lower central wing panel, it is necessary to proceed with some modifications.*

Ici, la construction préférée pour la conjonction *before* est le gérondif. D'autre part, la préposition *à* introduit l'argument 1. Dans la structure produite par l'analyseur, elle peut fort bien avoir été supprimée. *With* est contenu dans le cadre de valence de *proceed-V* pour la même position argumentaire, et est fabriqué par le générateur.

L'approche « fondée sur la connaissance » donne de très bons résultats par compréhension explicite, mais reste encore peu utilisable

L'approche *fondée sur la connaissance* (TAFC) a été défendue pendant près de vingt ans par Shank et son école comme la seule permettant de résoudre les problèmes de la TAO. Mais on ne réussissait pas à construire de maquette convaincante, tandis que les systèmes de TAFL progressaient et se diffusaient. Les sceptiques avaient beau jeu de souligner les difficultés théoriques et pratiques de l'entreprise. C'est seulement quand les chercheurs en IA ont abandonné leur quête irréaliste (ou prématurée) de solutions générales pour se tourner vers des objectifs réalistes mais limités, en développant des *systèmes experts,* ainsi que des outils permettant le développement de bases de connaissances non triviales, que cette idée devint réalisable.

Le premier prototype de système de TAO *fondé sur la connaissance,* KBMT-89 (Nirenburg *et al.* 1989), fut développé au CMT (Center for Machine Translation) de CMU (Carnegie Mellon University). Il utilisait une *ontologie* de son domaine (PC-XT et PC 5550 japonisé), le corpus étant constitué d'une vingtaine de pages tirées des manuels de ces PC. Son lexique représentait environ 1 200 termes, et son ontologie 1 600 *concepts.* Même dans ce cadre restreint, il apparut qu'on ne pouvait pas résoudre toutes les ambiguïtés automatiquement, et qu'il fallait donc une révision, ou un dialogue avec un humain. Cette dernière solution fut retenue, et le système fut muni d'un *augmenteur* (Brown 1989), posant des questions à un spécialiste ne connaissant que la langue source.

Le CMT a récemment trouvé une première application industrielle, et développé à partir de KBMT-89 le système KANT (Nyberg *et al.* 1992), utilisé pour traduire de la documentation d'équipements lourds chez Caterpillar. La différence essentielle avec KBMT-89 est que, pour supprimer toute interaction durant le processus de traduction, on exige que le texte d'entrée appartienne à un langage non ambigu strictement contrôlé, ce qui entraîne une interaction assez lourde au moment de la création.

Il y a actuellement environ 14 000 sens pour les termes généraux, et quelques centaines de termes spécialisés (non ambigus), ce qui est encore loin des tailles courantes dans les systèmes de TAFL. D'autre part, les traductions obtenues sont grammaticalement et stylistiquement très bonnes, mais seraient souvent rejetées par des réviseurs professionnels comme des paraphrases inexactes. Prenons deux exemples donnés par Nyberg et Mitamura.

Anglais source	Français cible	Allemand cible
In order to prevent a fire hazard, do not overload AC outlets.	Afin d'éviter tout risque d'incendie, ne jamais surcharger les prises CA.	Vermeiden Sie Feuergefahren, indem Sie die Netzanschlüsse nicht überlasten.
Traductions plus exactes (topic et focus ont été inversés en allemand, « tout » devrait venir de « any », et « jamais » de « never »)	*Afin d'éviter les risques d'incendie, ne pas surcharger les prises CA.*	*Um Feuergefahren zu vermeiden, überlasten Sie die Netzanschlüsse nicht.*
If the TV set has been dropped, a shock hazard may exist.	La chute du téléviseur peut provoquer un risque de choc électrique.	Wenn Sie das Fernsehgerät fallen lassen, kann die Gefahr eines Elektroschocks bestehen.
Traductions plus exactes (les relations sémantiques et temporelles ont été incorrectement traduites, « Sie » devrait venir de « you »)	*Si on a laissé tomber le téléviseur, il peut y avoir un risque de choc électrique.*	*Wenn man das Fernsehgerät hat fallen lassen, kann die Gefahr eines Elektroschocks bestehen.*

Ces traductions ne sont en fait pas meilleures que des traductions obtenues par des systèmes de TAFL spécialisés à des langages contrôlés analogues (Ducrot 1988), ou même à des sous-langages observés très restreints (Chandioux 1988), qui sont excellentes. Même si la qualité de la TAFC arrivait à dépasser celle de la TAFL sur ce genre de textes, on y gagnerait fort peu. On ne sait pas non plus si la TAFC peut produire des traductions brutes aussi bonnes que la TAFL sur des sous-langages observés (et non contrôlés) assez larges, – voir à cet égard les exemples de TAFL donnés plus haut.

La faisabilité technique de la TAFC a maintenant été établie. Pour l'instant, elle est cependant moins applicable que la TAFL, puisqu'elle impose un langage fortement contrôlé, et que, selon les mots (peut-être trop optimistes) de Nyberg et Mitamura eux-mêmes, « il est probable que la construction d'une base de connaissances sur le monde suffisante pour permettre la TAFC dans tout domaine de discours ne sera pas réalisée avant quelques années[7]. »

7. « It is probably the case that the implementation of a world knowledge base sufficient to support KBMT in any domain of discourse is some years from realization. »

Cependant, avec l'émergence des systèmes experts de grande taille, on peut espérer trouver dans le futur des entreprises ou des organismes qui auront développé de tels systèmes pour la CAO ou la CFAO, et qui auront besoin de traduire des volumes importants de textes portant sur les produits correspondants. On pourra alors développer des systèmes de TAFC sans avoir à construire (et à maintenir) une ontologie complète dans le seul but de traduire, ce qui coûte fort cher. L'ontologie pourra être obtenue à partir de la base de connaissances, ou même être réduite à une simple interface avec cette base, et il suffira de la coupler à un système de TAFL robuste (Gerber et Boitet 1985).

Conclusion sur la compréhension et la TAO

Il n'est absolument pas envisageable pour l'instant d'automatiser la traduction-recréation ni la traduction-localisation plus que par la mise à disposition d'outils d'aide au traducteur humain. Par contre, la *fonction traduisante* est automatisable dans les contextes de la traduction-diffusion et de la traduction-dépistage.

Dans certains cas, les systèmes de TAO « traduisants » *peuvent* comprendre (explicitement). Même quand cela est possible, un dialogue ou une révision semble rester nécessaire si l'on veut atteindre la qualité d'un spécialiste humain bilingue sans recourir à un langage strictement contrôlé. Dire que les systèmes de TAO *doivent* comprendre au sens fort serait se condamner à ne pas utiliser des systèmes de TAFL disponibles et... sans doute insurpassables dans certains cas.

L'utilisation de la *sémantique intrinsèque* des langues et des régularités des textes permet d'ores et déjà d'obtenir une qualité permettant de parler de *compréhension apparente directe,* suffisante pour la TAO du réviseur, et *a fortiori* pour la TAO du veilleur. Enfin, l'introduction d'un dialogue homme-machine approprié devrait permettre de construire des systèmes à *compréhension apparente indirecte* donnant des traductions de haute qualité sans révision de textes non contrôlés.

Aspects situationnels, scientifiques et ergonomiques de la TAFD

Opportunité de la TAFD

Motivations

Les recherches et développements en TAFD sont motivés par la limitation des paradigmes actuels, par l'importance croissante des langues nationales dans le contexte de l'internationalisation, et par de récents progrès méthodologiques et technologiques.

a. Limitation des paradigmes actuels

La TAFL est très bien adaptée à la TA du veilleur. Des systèmes portables de qualité tout à fait convenable pour l'usage visé commencent à se répandre au Japon, à des

prix abordables (environ 5 M¥ pour un système logiciel et matériel complet[8]). Par contre, elle est loin de pouvoir répondre à tous les besoins en traduction du réviseur. Outre le fait qu'elle demande évidemment autant de révisions que de langues cibles, elle reste trop chère pour des usages légers. En effet, selon les chiffres donnés par les producteurs de systèmes de TA (JEIDA 1989), la création *ex nihilo* d'un système opérationnel de TAFL coûte entre 200 et 300 hommes-années, avec des développeurs spécialisés, et l'adaptation d'un système de TAFL existant à un nouveau domaine et à une nouvelle typologie de textes coûte de l'ordre de 5 à 10 hommes-années. Un utilisateur n'a intérêt à s'équiper d'un tel système que s'il a à traduire de gros flux de textes homogènes et informatisés, comme des manuels d'utilisation ou de maintenance[9]. Adapter un système disponible à des besoins ponctuels n'est pas non plus une solution viable[10].

D'autre part, une condition essentielle de succès de ce type de TAO est de constituer une équipe de développement et de maintenance des linguiciels (dictionnaires, grammaires) qui soit en liaison constante avec l'équipe de révision, et si possible avec les auteurs des documents à traduire[11]. C'est ce qu'a réussi la PAHO (Pan American Health Association) (Vasconcellos et León 1988), avec ses systèmes ENGSPAN et SPANAM. On peut faire un parallèle avec les systèmes experts, qui peuvent être développés par des tierces parties, mais qui doivent ensuite être totalement maîtrisés par leurs utilisateurs, seuls à même de les faire évoluer de façon adéquate.

En ce qui concerne la TAFC, elle est totalement inapplicable en TAO du veilleur, et nous avons montré qu'elle était moins applicable que la TAFL en TAO du réviseur. Tant qu'il faudra construire les ontologies spécialement pour la traduction, elle restera aussi plus onéreuse.

b. Importance croissante des langues nationales et de l'internationalisation

De plus en plus, nous désirons rédiger dans notre langue, et transmettre nos textes à l'étranger, qu'il s'agisse de messages électroniques, de lettres, d'articles, de manuels techniques, voire de livres. Contrairement à ce que d'aucuns prédisaient il y a une cinquantaine d'années, l'internationalisation croissante ne s'est pas accompagnée d'une uniformisation linguistique vers l'anglais, mais au contraire d'un renforcement considérable de l'usage scientifique et technique des langues traditionnellement importantes de ce point de vue, et d'une promotion volontariste de bien d'autres, pour les amener au même niveau (malais-indonésien ou arabe, par exemple). À notre

8. Station Sparc, lecteur optique, système avec 80 000 termes, dictionnaire utilisateur, et options de personnalisation.
9. En prenant l'hypothèse d'un système coûtant 1 MF (400 KF de base et 600 KF de spécialisation au vocabulaire et au type de texte) et d'un amortissement sur 2 ans, il faut un flux de 10 000 p/an (en comptant 10 %/an de maintenance, 60 F/page de coût machine, et 100 F/page de révision, contre 150 F/page de traduction et 70 F/page de révision pour la méthode manuelle, soit 60 F/page de gain pour amortir 1,2 MF). À coût machine nul, il faudrait encore 5 000 p/an.
10. Ce serait comme réoutiller une usine pour produire quelques dizaines de voitures. En effet, sans compter la saisie optique ou manuelle, entraînant toujours un coût important de vérification, ni la maintenance, ni même l'achat du système de base, mais seulement sa spécialisation (600 KF) et les coûts de traduction et de révision, on arrive avec les hypothèses précédentes à 632, 680, 760 et 920 KF pour 200, 500, 1 000 et 2 000 pages, contre 44, 110, 220 et 440 KF pour la méthode classique manuelle, soit environ 14,5, 6, 3,5 et 2 fois plus, respectivement.
11. Dans le « contre-rapport ALPAC » du JEIDA (1989) comme au MTS-II à Munich en août 1989, Fujitsu reconnaissait clairement avoir fait une erreur en distribuant largement ATLAS-II : seules étaient en effet rentables les traductions effectuées chez Fujitsu, soit pour sa documentation, soit dans le cadre d'un contrat avec la CEE (Japinfo) concernant la veille technologique et ne demandant donc qu'une révision minimale.

sens, cette évolution ne fera que se renforcer, les langues étant, comme le notait le Professeur Hagège dans un article paru dans *Le Monde* début 1990, les « drapeaux des identités nationales ».

Il ne s'agit pas seulement de politique, mais d'efficacité. Dans les projets coopératifs européens (Esprit, Eureka), par exemple, la communication est gênée par la nécessité de lire et d'écrire en anglais. Pour la grande majorité des participants, lire en anglais pose des problèmes de compréhension et prend beaucoup de temps. Quant à écrire, si même c'est envisageable, le résultat est souvent difficile à comprendre, voire illisible.

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

c. Progrès méthodologiques et technologiques

L'idée de la TAFD date des années soixante (Kay 1973), et a été incorporée à plusieurs maquettes ou prototypes dans les années soixante-dix et quatre-vingts (Melby, Smith et Peterson 1980 ; Tomita 1984 et 1985 : Hutchins 1986 ; Chandler, Holden, Horsfall *et al.* 1987 ; Weaver 1988 ; Wood et Chandler 1988 ; Wood 1989). Si ces travaux n'ont pas donné lieu à des systèmes utilisables en pratique, c'est que les dialogues devaient être conduits par des spécialistes[12], que la couverture linguistique était trop limitée, et que l'on ne disposait pas encore d'environnements interactifs conviviaux.

La méthodologie s'est affinée ces dernières années. Tout d'abord, l'utilisateur envisagé n'est plus un spécialiste, mais un rédacteur, ou plutôt un *auteur* (Boitet 1989a et 1990b ; Sadler 1989 ; Blanchon 1990 ; Huang 1990 ; Maruyama *et al.* 1990 ; Somers, Tsujii et Jones 1990). Nous préférons ce dernier terme. D'un côté, en effet, *auteur* est moins restrictif que *rédacteur* : un auteur est quelqu'un qui veut créer un texte, et peut le faire en l'écrivant, en le dictant, ou encore en le construisant interactivement. De l'autre, *auteur* est plus restrictif que *rédacteur, locuteur,* ou *commentateur*, car *auteur* désigne quelqu'un qui désire créer un produit final *propre*, alors que les autres termes peuvent renvoyer à des personnes désirant seulement produire un message écrit ou parlé de façon *spontanée*, en vue d'une communication immédiate, et non disposées à conduire un dialogue éventuellement lourd pour rendre leur message *propre*[13].

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

12. ITS (Melby *et al.* 1980) demandait même *plusieurs* intervenants, un linguiste spécialiste du système pour l'analyse, et un bilingue pour chaque langue cible.
13. « Propre » signifie ici conforme à une certaine grammaire (permettant éventuellement des constructions incorrectes, pourvu qu'elles soient attestées) *et* ne comportant pas de parties « réflexives », ou « automodificatrices », si fréquentes dans la parole et même l'écriture spontanée (au moins manuscrite), comme des hésitations, des faux départs, des reprises, des répétitions, des corrections, des abréviations arbitraires (apocopes), etc.

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

Situations traductionnelles pour la TAFD

a. Critères de choix de l'approche par dialogue

Nous proposons quatre critères pour choisir la TAFD :

- la qualité visée doit être élevée, et la révision impossible ou très coûteuse ;
- le contexte doit être fortement multilingue (1—> n, comme pour la dissémination de documentation technique, ou n <—> n, comme dans des projets internationaux) ;
- l'entrée ou le domaine ne doivent pas être trop contraints ou contrôlés (sinon, mieux vaut utiliser la TAFL ou la TAFC) ;
- les utilisateurs doivent être prêts à participer à des dialogues de normalisation et de désambiguïsation.

Dans toute situation, il faut de plus pouvoir rendre les dialogues acceptables, en les laissant à l'initiative de l'utilisateur, en lui fournissant des moyens de les contrôler et de les réduire (en jouant sur des paramètres, en insérant directement des marques de désambiguïsation, etc.), et en lui laissant si possible le choix entre plusieurs média.

b. Situations traductionnelles adaptées à la TAFD

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

- la traduction de volumes relativement faibles de documentation technique en plusieurs langues, typiquement 5 000 à 8 000 pages à distribuer sur un DON[14], par exemple dans les 9 langues de la CEE (et peut-être dans d'autres, comme le russe, l'arabe, le japonais, ou le chinois).
- la diffusion d'information dans plusieurs langues (sur la circulation, sur la météo, ou dans des congrès, des manifestations sportives, des situations d'urgence...), qui demande une sortie orale aussi bien qu'écrite ;
- l'échange télématique de notes et de documents de travail dans des projets internationaux.

Comme l'état de l'art en reconnaissance de parole ne permet pas de traiter à la fois un grand vocabulaire, de la parole continue, et un locuteur arbitraire, il semble n'y avoir que très peu de situations favorables à la TAFD avec entrée orale :

- la production de commentaires ou de résumés à partir de scènes visuelles et auditives (par exemple, pour le sous-titrage d'émissions de télévision en langues étrangères), ou encore la préparation de la partie sonore de bandes vidéo en plusieurs langues ;

14. Disque optique numérique (CD-ROM), contenant près de 600 M d'octets.

- l'interprétation de dialogues bilingues très contraints, tels que les appels téléphoniques de politesse entre parents d'enfants échangés entre familles pour apprendre les langues, ou l'assistance téléphonique à des voyageurs étrangers (consultation médicale, réservation...). Ici, le dialogue doit être le plus réduit possible, et une combinaison entre TAFD et TAFC (analogue à l'architecture finale de KBMT-89) semblerait indiquée.

Il y a enfin beaucoup de situations intéressantes où le message source n'est créé que pour vérifier le contenu du (ou des) message cible, résultant d'une négociation avec un expert. C'est, par exemple, le cas de lettres officielles ou formelles, qui ont des structures très différentes dans différentes cultures. Somers, Tsujii et Jones (1990) ont proposé le terme de *traduction sans texte source*, mais il est peut-être plus exact de parler de génération multilingue de textes que de TAFD.

Aspects scientifiques

La controverse entre praticiens et théoriciens

Comme nous le disions en commençant, les avis sur la traduction automatique (TA) sont souvent extrêmes, et la controverse est souvent vive. Certains ne la conçoivent que comme l'expérimentation scientifique de leurs théories ou de leurs formalismes favoris. À COLING-90, par exemple, on ne comptait plus les titres du genre *le formalisme XYZG et son application à la TA*, alors que cette application était tout au plus mentale et hypothétique. Dans un autre registre, M. Gross a soutenu pendant des décennies qu'on ne pouvait espérer faire de la TA sans avoir « mis la langue à plat ». Outre que cette ambition est sans doute illusoire, dans son principe même, la langue étant essentiellement productive, la réalité a démontré par l'absurde la fausseté de ce point de vue, puisqu'en fin de compte il y a des systèmes qui traduisent réellement, et bien, mais justement parce qu'ils sont limités à des sous-langages.

D'autres voient la TA comme une entreprise essentiellement technologique et utilitaire. A. Colmerauer, qui fit de la TA à l'Université de Montréal, a parlé d'un « gigantesque bricolage ». M. Nagao insiste aussi souvent sur le fait que les théories n'apportent qu'un cadre incomplet, et que l'essentiel du travail linguistique consiste à décrire une énorme quantité d'exceptions. Y. Wilks, chercheur en IA et auteur d'une expertise sur Systran, soutint même, au séminaire « Sémantique formelle et linguistique computationnelle » (Lugano 1988), que « il n'y a pas de théorie linguistique assez mauvaise pour ne pas faire de TA avec » !

D'autre part, vouloir développer une théorie élégante et l'utiliser comme fondement exclusif d'un système opérationnel mène à l'échec sur les deux fronts, comme cela a été démontré avec éclat par le projet Eurotra, et par plusieurs autres projets centrés sur les grammaires d'unification : la théorie ne progresse pas d'un iota, et on n'obtient pas de système utilisable.

Comme B. Vauquois, M. Kay, S. Nirenburg, M. Nagao, Yu. Apresyan, et bien d'autres, nous préférons dire clairement que la TA est essentiellement une entreprise technologique, bien qu'elle entretienne avec les diverses sciences qui la sous-tendent (linguistique, informatique, ergonomie) des relations tout à fait analogues à celles du

génie civil avec la physique. L'incorporation d'idées théoriques (mais non de théories entières) peut mener à des progrès tangibles, mais seulement de façon incrémentale, et souvent après un recul initial des performances. En retour, la pratique amène parfois à poser des questions théoriques intéressantes.

Des progrès techniques peuvent venir de la théorie

Citons brièvement quelques exemples d'idées théoriques qui ont fait progresser la TA. Il y a d'abord eu l'incorporation de formalismes chomskyiens comme les automates d'états finis, les grammaires hors contexte, et les grammaires transformationnelles, qui ont permis de travailler sur des représentations arborescentes des phrases au lieu de se limiter aux chaînes. Ces formalismes ont été immédiatement étendus (calculs d'attributs, primitives de contrôle, etc.) pour devenir utilisables en pratique mais ces extensions n'ont été reprises dans la théorie que 20 ans plus tard (GPSG).

Il y a ensuite l'usage de structures de dépendances, adaptées de Tesnière, de l'école de Prague, et de l'école russe, et combinées avec des idées de la logique formelle, ce qui a mené à des structures comportant prédicats, arguments et circonstants, les circonstants (et plus rarement les arguments) portant des *relations sémantiques*, ou *cas profonds*.

L'introduction d'unités lexicales (familles dérivationnelles) comme éléments lexicaux de base pour le transfert et la génération se révéla très fructueuse, mais ne fut possible qu'en simplifiant notablement la théorie de Mel'čuk, par réduction de l'ensemble des fonctions lexico-sémantiques aux principales dérivations productives.

Enfin, il faut admettre que la technique fait feu de tout bois, et que les systèmes qui tournent se rattachent à *plusieurs* théories à la fois. S. Nirenburg a trouvé un terme qui exprime très bien ce que nous venons de dire : un bon système de TA doit être capable d'intégrer progressivement de nombreuses *microthéories*.

Y a-t-il des idées théoriques inutilisables en TAFL ou TAFC, et qui pourraient bénéficier à la TAFD ? Nous le pensons. Un premier exemple serait l'utilisation effective de la notion zembienne de *statut*. La *triade statutaire* est la décomposition d'une proposition en thème, rhème et phème (Zemb 1982)[15], et est très importante pour la traduction (choix des articles, portée et place de la négation...). Mais elle est pratiquement impossible à calculer automatiquement, même s'il y a des critères formels pour certains cas précis (subordonnées négatives en allemand, *ordre communicatif* des circonstants dans les langues slaves...). Pour déterminer le statut d'un terme de la proposition, il convient de poser des questions à son auteur. Et c'est justement ce que permet la TAFD.

Par exemple, pour traduire la phrase suivante (adaptée de Chomsky), on peut demander si Jean est un laveur de voitures. Si oui, *les voitures* est rhématique (première traduction).

Jean ne lave pas les voitures — *Hans wäscht keine Wagen*
Hans wäscht die Wagen nicht

15. Le rhème est ce qu'on dit du thème, et le phème la modalité (kantienne) de cette assertion. Le thème persiste dans l'existence quand le phème varie : quand « Hans wäscht die Wagen nicht », il y a toujours des voitures à laver !

D'autres exemples concernent la détermination de l'aspect et de la modalité, toujours au moyen de questions. Voici un exemple dû à Tomita (Blanc et Boitet 1990) :

Le courrier est arrivé ce matin	*The mail arrived this morning* *The mail has arrived this morning*

La technologie suscite des questions théoriques

Dans la majorité des situations adaptées à la TAFD, il faut un système de couverture lexicale et grammaticale très large. Cela amène à poser aux théoriciens des questions importantes :

- Sachant qu'on n'obtient de bons résultats que sur des langages restreints, *comment construire une base de connaissances linguistiques utilisable comme une union de sous-langages ?* Est-il possible de séparer les aspects grammaticaux et lexicaux ?
- *Comment atteindre la large couverture nécessaire ?* Typiquement, un système de TAFL contient de 3.10^4 à 3.10^5 termes, en 2 langues. Le cas de METEO (3.10^3) est atypique, à cause de son domaine très restreint. Mais un système de TAFD visant le grand public et non restreint à un domaine particulier demandera de 3.10^5 à 3.10^6 termes, en plusieurs langues !
- Dans des situations fortement multilingues, l'approche par interlingua est séduisante. Mais, *comment surmonter les difficultés d'ingénierie rencontrées dans la construction d'un grand lexique interlingue* par les récents projets japonais ATLAS, PIVOT, EDR, et CICC ?
- Il est crucial que des non-spécialistes puissent facilement comprendre les questions du système, éventuellement lui demander les raisons de certaines questions, et comprendre ses réponses. Une question importante (et nouvelle) est donc de trouver *comment rendre la base de connaissances linguistiques d'un système de TAFD accessible à un utilisateur naïf.*

Aspects ergonomiques

Les aspects ergonomiques sont bien sûr importants en TAO classique. En TA du veilleur ou de l'auteur, il y a une notion de qualité *apparente*, liée à la présentation. La même traduction paraît bien meilleure si elle est formatée comme un texte que si elle est présentée phrase par phrase, ou pire en colonne. De même, on accepte plus volontiers des choix portant sur de petits groupes de mots que sur des phrases complètes. Les constructeurs de systèmes ont beaucoup travaillé sur les éditeurs bilingues, sur les outils de paramétrage, et sur les aides à la création de dictionnaires.

En TAFD, l'ergonomie est un aspect absolument crucial, et les choix ergonomiques influencent directement l'architecture de tout le système. Les choix principaux sont les suivants :

- Le système doit-il fonctionner en temps réel, ou l'asynchronie est-elle préférable ?
- Doit-il tourner sur des ordinateurs personnels bon marché, ou sur des stations de

travail ? Une architecture avec serveur est-elle possible ? Si oui, pourrait-on simplement connecter un PC au minitel ?

- Comment les dialogues doivent-ils être organisés ? Est-il nécessaire et/ou possible de les conduire dans un environnement multimédia ? Plus précisément, l'utilisation de synthèse de parole peut-elle améliorer l'efficacité et la convivialité des dialogues de désambiguïsation ?

Architecture linguistique et voies intermédiaires nouvelles

Le cadre nouveau de la TAFD peut amener à rechercher de nouvelles architectures linguistiques. Il ne s'agira sans doute pas de solutions radicalement nouvelles, mais, comme souvent dans un domaine technique, de nouveaux compromis, de voies intermédiaires nouvelles, avec ça et là des innovations intéressantes.

Transfert multiniveau à acceptions, propriétés et relations interlingues

Les systèmes de TAO modernes sont fondés sur un *transfert sémantique*, le *passage par un interlingua,* ou un t*ransfert multiniveau. Transfert sémantique* est un terme introduit par les Japonais pour désigner exactement l'approche du CETA entre 1960 et 1970, que B. Vauquois appelait le *pivot hybride :* la structure interface source fournie au transfert contient des éléments lexicaux de la langue source, et des propriétés et relations interlingues (traits sémantiques, cas profonds, nombre, aspect, modalité, détermination abstraite...). Dans un *interlingua,* les éléments lexicaux sont de plus interlingues (les auteurs des systèmes actuels parlent de *concepts,* mais il n'y a pas toujours d'ontologie).

Le transfert (sémantique) multiniveau, au sens de Vauquois (Vauquois et Boitet 1985 ; Vauquois 1988) diffère du transfert sémantique en ce que, outre les attributs et relations interlingues, on garde des attributs et des relations spécifiques à la langue source (classe syntagmatique, genre, nombre, détermination, temps, mode, fonction syntaxique...)

En TAFD, nous proposons de rajouter aux représentations multiniveaux un niveau lexical, celui des *acceptions interlingues,* sans aller jusqu'à introduire des concepts, puisqu'il faudrait alors construire une ontologie. La base lexicale multilingue sous-jacente (BDLM) contiendra alors un dictionnaire monolingue pour chaque langue traitée par le système, et un dictionnaire interlingue pour les acceptions interlingues. Chaque acception interlingue a une image dans chaque dictionnaire monolingue, avec une définition appropriée dans la langue correspondante, utilisée lors de la désambiguïsation interactive du sens[16].

Remarquons aussi que *interlingue* ne signifie pas *indépendant des langues.* Par exemple, si le système travaille avec le français, l'anglais et le russe, il y aura une seule acception pour *mur* en tant qu'objet concret. Dès qu'on ajoutera l'allemand ou

16. On peut aussi *expliquer* à l'auteur pourquoi une telle question est posée, et même montrer les mots en question dans les autres langues. L'introduction d'aspects d'auto-apprentissage dans ce genre de système les rendrait plus acceptables par les utilisateurs potentiels.

l'italien, il faudra ajouter les raffinements *mur vu de l'extérieur (Mauer, muro)* et *mur vu de l'intérieur (Wand, parete).*

Langage guidé

Séparation lexique/grammaire

La notion de « sous-langage » a été introduite et étudiée par le linguiste R. Kittredge (Richardson 1985 ; Vauquois et Chappuy 1985), à la suite de son expérience en TAO (il fut directeur du groupe TAUM de l'Université de Montréal au début des années 70). Kittredge a donné un certain nombre de critères, essentiellement lexicaux et syntaxiques, pour évaluer la difficulté d'un sous-langage pour les techniques de TAO du réviseur, et pour déterminer si l'approche dite de « deuxième génération avec sous-langage » (ou *sous-optimisation*) était prometteuse, et les a appliqués à un certain nombre de types de textes.

Son analyse, extrêmement complète, distingue les aspects lexicaux et grammaticaux, la liaison plus ou moins forte avec un domaine sémantique clos, et la possibilité d'écrire une grammaire textuelle dépassant le niveau de l'énoncé. Il introduit la notion formelle de *clôture lexicale*, qui signifie en gros que le nombre de nouveaux termes rencontrés dans une nouvelle page diminue rapidement et tend vers zéro ou une valeur très faible quand le nombre de pages augmente.

Mais il ne propose aucune notion formelle analogue pour l'aspect grammatical : un sous-langage est défini comme l'ensemble des phrases (ou des énoncés) produisibles dans des conditions fixées (par exemple, bulletins météo, appels d'offres du Secrétariat d'État, manuels de maintenance de tel avion, etc.), sans qu'on sache comment choisir un échantillon convenable et comment généraliser autrement qu'intuitivement. De plus, le terme choisi est gênant : Kittredge fait lui-même remarquer qu'un sous-language d'une langue n'est pas un sous-ensemble de cette langue (cela est dû au fait que le *français* signifie en général le français standard des manuels, et pas l'union de tous ses jargons et parlers régionaux).

D'autre part, si l'usage de la TA se répand grâce à la TAFD, il sera impossible de produire et de maintenir une collection très variée de bases lexicales et grammaticales de grande taille correspondant à des sous-langages au sens de Kittredge, une pour chaque type d'utilisation. À terme, il faudra donc un dictionnaire total aussi exhaustif que possible, et cela ne sert à rien d'utiliser la notion de *clôture lexicale* pour le limiter. La même chose est vraie de la grammaire.

Pour réduire le problème (diviser pour régner !), il y a une voie intermédiaire, qui consiste à séparer les deux aspects, puis à définir chacun d'eux en deux temps, d'abord de façon grossière à l'aide d'un formalisme symbolique simple, puis de façon plus fine en ajoutant des paramètres numériques.

Un type de texte donné pourra alors être défini comme un ensemble de poids relatifs à des arcs et à des nœuds d'une BDLM structurée en réseau sémantique, son *profil lexical,* et comme un style d'énoncés ou un genre de texte vérifiant certaines contraintes numériques.

Préférences lexicales

La BDLM d'un système de TAFD doit contenir des relations entre acceptions qui permettent d'en extraire des thesaurus, en particulier la synonymie, la quasi-synonymie et l'hypéronymie, et des relations analogues aux relations lexico-sémantiques entre termes (comme entité —> qualité, action —> argument 1 de l'action...).

Les poids attachés aux acceptions, aux termes et aux relations entre eux constituent ce qu'on pourrait appeler un *profil lexical.* Au fur et à mesure que le temps passe, le système de TAFD peut les faire varier en fonction de l'interaction, ce qui permet un certain réglage automatique, et la définition de nouveaux profils lexicaux reflétant les préférences lexicales courantes. Les poids sur les termes reflètent leur *degré de pertinence* par rapport à la tâche en cours, et peuvent être utilisés pour indiquer à l'auteur le terme préféré parmi un ensemble de (quasi-)synonymes (comme par exemple avion, appareil, aéronef). Associés aux poids sur les acceptions et sur les relations entre termes et acceptions, ils peuvent aussi être utilisés pour lever des ambiguïtés de sens, comme cela a été montré dans Chandioux (1988). En TAFD, cela peut servir à présélectionner le sens le plus probable.

Remarquons que, dans le contexte d'un système de *TAFD pour tous,* la base lexicale devra contenir une très grande variété de termes, même incorrects ou douteux, alors que les bases de données terminologiques ne contiennent d'habitude que des termes normalisés ou recommandés.

Types de textes : « styles d'énoncés » et « genres de textes »

En ce qui concerne l'aspect grammatical, on peut proposer de diviser encore le problème en deux, en définissant des *styles d'énoncés* pour les phrases et autres énoncés traduisibles individuellement (titres, éléments homogènes de longues énumérations...) et des *genres de textes* pour les textes plus longs[17]. Pour cela, nous supposons que nous avons recensé toutes les constructions d'une langue, y compris les constructions rares ou n'apparaissant que dans des types de textes très techniques (par exemple : « Mettre interrupteur sécurité train avant sur OFF »), et que nous les avons représentées au moyen d'un très grand ensemble R de règles déclaratives. Tout formalisme déclaratif simple convient[18].

Un *style d'énoncé* est alors un sous-ensemble de R vérifiant certaines restrictions numériques (par exemple, sur le degré d'imbrication, d'ellipse, ou de coordination). Par exemple, M1 pourrait être le style des phrases simples sans sujet (« permet de sauver sur disque une copie de votre fichier »), fréquentes dans les documents techniques, et M2 le style des phrases explicatives simples.

En ce qui concerne les *genres de textes,* il est souhaitable qu'on puisse dans le

17. Les termes de « microlangage » et de « sous-langage » proposés dans Boitet (*1990b*), se sont révélés surchargés et anti-intuitifs.

18. Par exemple, les grammaires hors contexte (CFG), avec ou sans attributs, les TAG, les STCG (Vauquois et Chappuy 1985 ; Zaharin 1986 ; Boitet et Zaharin 1988). Pour le moment, nous utilisons ROBRA dans LIDIA-1, malgré son caractère procédural, et testons dans chaque règle transformationnelle si le style d'énoncé attendu est l'un des styles contenant cette règle.

futur les prendre en compte dans des éditeurs de documents structurés fondés sur SGML, qui peuvent traiter des textes beaucoup plus longs que les outils linguistiques actuels, le plus souvent limités à la phrase ou au paragraphe. On peut alors proposer de définir un genre de texte comme une expression algébrique sur les styles d'énoncés et les genres de textes. Par exemple, le genre de texte S1 des paragraphes commençant par une phrase de style M1 suivie par une suite (éventuellement vide) de phrases de style M2 peut être défini par une simple expression régulière :

```
<S1>          =  M1  M2*
```

Pour décrire le genre de textes d'un *document* commençant par un titre (microlangage M3), et se poursuivant par une liste non vide de paragraphes (microlangages M2 et/ou M4) et/ou de sections de même structure qu'un document, on peut de même écrire[19] :

```
<Document>    = <Title> <Content>
<Title>       = M3
<Content>     = (<Paragraph> | <Document>) +
<Paragraph>   = (M2 | M4) +
```

Il faudra plus de recherche pour trouver comment guider les auteurs dans la sélection d'un style d'énoncés ou d'un genre de textes particulier, de façon à ce que la critique textuelle, la standardisation et la clarification puissent être menées avec efficacité.

Accessibilité des connaissances

Dans la plupart des systèmes de TAO, les informations linguistiques sont très détaillées, et codées de façon compréhensible uniquement par des spécialistes. Dans certains, qui se présentent plutôt comme des aides au traducteur humain (Weidner, ALPS), on a au contraire cherché à n'utiliser que des informations très simplifiées, pour que des utilisateurs naïfs puissent coder eux-mêmes les dictionnaires. Le résultat a simplement été que les traductions étaient trop mauvaises pour servir de base à une révision[20].

En TAFD, il nous semble que l'information ne doit pas être simpliste, mais peut-être moins fouillée que dans le cas de la TAFL. Par exemple, il est sans doute inutile d'avoir un système trop riche de codes sémantiques. Une hiérarchie à 3 ou 4 niveaux, avec au plus une dizaine de codes par niveau, est sans doute le maximum si l'on veut que des utilisateurs puissent compléter les dictionnaires, qui, même très grands, ne seront jamais complets.

Un second aspect, très délicat et intéressant, est de trouver comment exprimer les notions linguistiques obscures pour le commun des mortels d'une façon compréhensible. Pour certaines, comme l'aspect (voir plus haut l'exemple du courrier qui

19. Un symbole entre crochets désigne un genre de texte, « | » l'alternation, « * » et « + » la répétition. On pourrait ajouter des conditions sur des attributs associés aux symboles principaux.
20. En 1985, le Bureau des Traductions d'Ottawa mena une étude consistant à comparer la productivité des traducteurs humains utilisant le système Weidner sans, puis avec l'option de TA, en leur faisant traduire le même livre sur l'odontologie à 3 mois de distance. Dans le second cas, la productivité baissa de 40 % !

« est arrivé »), il faut sans doute utiliser des exemples, ou de la reformulation, et éviter de parler de la notion elle-même.

Enfin, il faut arriver à organiser le système de façon à ce que l'utilisateur, même monolingue, puisse contrôler ce que produit le système dans les langues cibles. Nous proposons pour cela un mécanisme de *rétrotraduction* (voir *infra*).

Annotations et préédition indirecte

L'idée de base de la TAFD est de remplacer la postédition par une préédition indirecte. Cela signifie que le texte est enrichi, normalement indirectement, grâce à l'interaction avec l'auteur. Mais des utilisateurs expérimentés doivent pouvoir faire une partie de cette préédition directement, pour éviter de longs dialogues. C'est pourquoi il faut représenter un texte, avec son système d'écriture, sa structure logique, ses marques de désambiguïsation, et ses résultats d'analyse, par une chaîne de caractères *portable* et *lisible,* dans l'esprit de la TEI (*Text Encoding Initiative*).

Une situation concrète pour la TAFD et la maquette LIDIA-1

Pour l'instant, la TAFD est encore en phase de recherche préliminaire. Au GETA, nous avons lancé autour de ce concept le projet LIDIA (Large Internationalisation des Documents par Interaction avec leurs Auteurs). Pour commencer, nous nous limitons à une situation particulière, où un rédacteur monolingue produit de la documentation technique à traduire en plusieurs langues.

Dans une première étape, nous construisons une maquette de petite taille, LIDIA-1, dont l'objectif est uniquement de nous permettre d'expérimenter une architecture nouvelle, et d'attaquer un certain nombre de problèmes linguistiques, informatiques, et ergonomiques. Cependant, notre souci constant est d'effectuer des choix cohérents avec l'énormité des bases lexicales et grammaticales qu'un système grand public devrait offrir pour être viable.

Quelques choix techniques

Situation multicible

Pour LIDIA-1, on suppose qu'un ingénieur français crée de la documentation technique sous la forme d'une pile HyperCard, sur un Macintosh de puissance moyenne, et aide le système à la traduire en anglais, allemand et russe. Nous avons choisi une architecture distribuée (station de travail de l'auteur sur Macintosh et serveur de TAO sur un mini IBM-4361). Le choix des langues est uniquement dû aux compétences disponibles, et le choix d'une situation monosource et multicible permet évidemment de réduire le coût de l'opération en n'ayant qu'une analyse à construire ! Il y a des arguments moins contingents pour les deux autres choix.

Hypertexte

Avec l'arrivée d'HyperCard™, les hypertextes sont sortis des laboratoires. Pour un prix très modique, n'importe qui peut réaliser des documentations vivantes, des animations, etc., avec image et son en prime. Certaines documentations de voitures (Renault, Peugeot) sont diffusées sur DON sous HyperCard (8 000 pages de documentation, en 10 langues, avec figures et éventuellement messages oraux, sous forme phonétique codée, 30 à 40 fois plus compacte que le signal).

Du point de vue ergonomique, l'hypertexte privilégie l'interaction. Par conséquent, on peut penser qu'un rédacteur sera plus prêt à accepter une interaction linguistique sous hypertexte que sous traitement de texte. Dans le premier cas, on reste dans la logique de l'outil. Dans le second, il faut changer ses habitudes, ce qui est souvent fort difficile.

Du point de vue linguistique, les parties textuelles sont bien isolées, et connexes, au contraire des traitements de texte ou des logiciels de PAO, où l'on trouve un mélange de codes de formatage, de texte, de figures, de formules, le tout parfois présenté de façon non connexe (tableaux avec tabulations...). De plus, les fragments de texte d'une pile HyperCard typique (champs et boutons) sont petits, et grammaticalement très homogènes : on peut parler de *types de fragments* beaucoup mieux que de *types de documents.* Par exemple, un champ donné peut contenir, dans chaque carte, un titre avec verbe à l'infinitif (« Naviguer dans HyperCard »), un élément de menu fonctionnant comme un nom dans d'autres contextes (« Quitter » – « ...cliquer sur Quitter »), un ordre à l'infinitif (« Prendre la disquette »), une explication simple sans sujet pour la première phrase (« vous permet de quitter l'application et... »), ou encore un titre simple sans verbe (« Rotation à gauche »), etc.

Le concept de TAFD n'impose pas l'utilisation d'hypertexte. On pourrait également le mettre en œuvre dans le cadre de *texteurs* ou de *documenteurs* plus classiques. Certains, comme WinText™, contiennent déjà des marques indiquant, outre la fonte, le corps et le relief, la langue naturelle de segments de texte. En ajoutant le type de fragment, et en développant une interface analogue, on pourrait les étendre pour y introduire les fonctions précédentes. Malgré tout, il faudrait se limiter à des texteurs ou documenteurs programmables (comme Interleaf™), ou modifier les codes sources. Le même problème se pose d'ailleurs aux concepteurs de *postes du traducteur.*

Architecture distribuée et traitements asynchrones

Nous désirons que la station de rédaction soit un micro-ordinateur convivial et largement diffusé, d'où le choix du Macintosh sous HyperCard. Nous désirons aussi réutiliser la puissance de notre générateur de systèmes de TAO (Ariane-G5), pour écrire les parties *lourdes* du traitement linguistique. Or, il ne tourne que sur mini IBM (Eurolang l'a en partie porté sur SUN en 1992), mais pas sur Macintosh. Quand bien même il le serait, son exécution, même en tâche de fond, serait trop lente sur ce type de matériel, et augmenterait inévitablement les temps de réponse.

D'autre part, l'exemple du système CRITIQUE (Richardson et Braden-Harder 1988) a montré qu'un traitement asynchrone convenablement organisé pouvait donner

un système très convivial. Nous avons donc opté pour un traitement distribué entre la station de rédaction et un serveur de TA, et un fonctionnement asynchrone, pour ne pas pénaliser le rédacteur par des interruptions forcées ou des attentes interminables. Tout doit rester sous son contrôle.

Enfin, ce type d'organisation devient de plus en plus réaliste, avec la disponibilité de réseaux télématiques puissants et relativement peu coûteux. Il est tout à fait possible d'imaginer un système de TAFD où le micro se connecte au serveur à des intervalles réglables par l'utilisateur, et où la connexion ne dure que quelques secondes, tout comme un utilitaire de messagerie.

Aspects informatiques de la maquette LIDIA-1

Intégration dans HyperCard et piles « traduisibles »

a. HyperCard

Il y a cinq sortes d'objets en HyperCard, les boutons, les champs, les cartes, les fonds et les piles. Les boutons sont des zones actives de l'écran, qui provoquent des actions quand ils sont *cliqués*. Les champs peuvent contenir du texte éditable, et les boutons du texte non éditable.

Une carte a ses propres boutons et champs, et un fond qui a à son tour des boutons et des champs. Un fond peut être partagé par plusieurs cartes. La carte recouvre son fond (les deux ont la même taille). Des dessins peuvent être dessinés sur les cartes et sur les fonds.

Les cartes sont regroupées en piles, chaque pile étant un fichier Macintosh. Une pile peut avoir plusieurs fonds. L'utilisateur communique avec HyperCard en agissant sur la carte affichée à l'écran, le mode exact d'interaction étant déterminé par un jeu de préférences. Nous avons ajouté une nouvelle préférence (une case à cocher) pour lancer ou arrêter LIDIA-1.

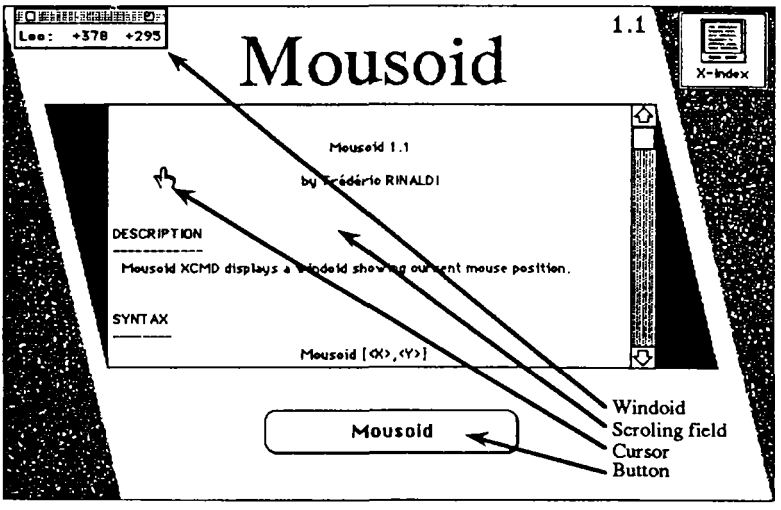


FIGURE 1 : Exemple de carte HyperCard.

b. Contraintes à respecter pour qu'une pile soit traduisible

Nous dirons qu'une pile est organisée de façon *traduisible* si l'on n'a pas à traduire les scripts[21], mais seulement les noms des boutons et les contenus textuels des champs[22].

Les scripts doivent donc être écrits indépendamment des langues d'où les restrictions suivantes :

- les références à des boutons ne doivent jamais être leurs noms, mais leurs numéros, invariants en traduction ;
- les messages ne doivent jamais être contenus dans les scripts, mais toujours pris dans des champs normaux[23], invisibles dans la version finale de la pile ;
- le texte contenu dans les dessins ne doit pas être traduit ;
- toute version personnalisée de la barre de menus doit être traduite à la main.

c. Scénario

Si l'auteur coche la case de préférence HyperCard LIDIA, le Macintsoh se connecte périodiquement au serveur de TAO. L'utilisateur travaille normalement sur sa pile. Quand il décide que certains objets (champs, boutons, cartes, ou toute la pile) sont prêts à être traduits, il active l'outil *traduction* dans la palette LIDIA, sélectionne ce ou ces objets, et continue à travailler pendant que les processus liés à la traduction s'exécutent.

L'état de traduction de tout objet peut être contrôlé grâce à un *témoin d'état (status watcher)*. Quand l'intervention de l'utilisateur est requise, LIDIA envoie un signal (paramétrable), exactement comme une tâche de fond comme PrintMonitor™ ou Eudora©. L'utilisateur est libre d'interagir tout de suite ou plus tard.

Pour interagir avec un objet, on double-clique simplement sur son témoin d'état et on active l'item approprié dans le menu déroulant qui apparaît. Le mode interaction peut être quitté à tout moment. Les objets changent d'état lors de l'interaction et sont repris en charge par LIDIA pour poursuivre le traitement si nécessaire.

Organisation des traitements linguistiques

Expliquons brièvement les traitements principaux illustrés dans la figure ci-dessous.

1. Après l'étape de standardisation,
- tous les champs et boutons doivent porter un type de texte[24], permettant de

21. Chaque objet d'HyperCard a un *script*, éventuellement vide, écrit dans le langage de programmation HyperTalk. Un script est une collection de *handlers*, chacun de la forme « on<message>do<suite d'instructions HyperTalk>end ». Un handler est invoqué quand son <message> (un clic de souris, par exemple) est reçu par l'objet contenu dans le script.
22. On a fait ici le choix de traduire une pile complètement, en autant de piles qu'il y a de langues cibles. Une autre possibilité serait de rendre la pile multilingue en créant une copie de chaque conteneur de texte pour chaque langue cible.
23. Cependant, elles peuvent contenir des variables : « Où est le fichier &1 ? » doit certainement être traduit.
24. Dans le cas de textes « incomplets », par exemple si le sujet de la première phrase est contenu dans un autre champ (comme dans des tables contenant des noms de commandes et leurs explications), ce module demande aussi comment construire le texte complet.

contrôler le correcteur stylistique, puis de guider l'analyse (d'où un module de *catégorisation textuelle*) ;
- les textes doivent être orthographiés correctement, et être conformes aux paramètres stylistiques associés au type de leur conteneur (nous sommes en train d'intégrer les *correcteurs grammaticaux et stylistiques* gracieusement fournis par Machina Sapiens) ;
- les syntagmes figés se comportant de façon spéciale (comme l'item de menu Cacher les bulles d'aide dans « Cacher les bulles d'aide arrête l'aide par bulles ») doivent être marqués (l'utilisateur doit aider le *module de syntagmes figés spéciaux* à établir la liste des syntagmes figés spéciaux de la pile)[25] ;
- les préférences terminologiques (voir *supra*) doivent avoir été plus ou moins imposées par le *module de préférence lexicale,* selon le degré de normalisation terminologique souhaité.

2. Le texte standardisé est alors analysé sur le serveur. La *mmc-structure source* produite (Multisolution, Multiniveau et Concrète) est transformée en une forme portable et lisible (directement par Lisp... et par les développeurs) et envoyée au Macintosh.

3. La mmc-structure source est utilisée pour produire le dialogue de désambiguïsation sur le Macintosh. Le processus de désambiguïsation la transforme en une *umc-structure source* non ambiguë (Unisolution, Multiniveau et Concrète) correspondant à l'analyse choisie par l'auteur[26].

4. Cette umc-structure source est alors « abstraite », ou « réduite », à une *uma-structure source* (Unisolution, Multiniveau et Abstraite).

5. À partir de la uma-structure source, le système Ariane-G5 produit les *gma-structures cibles* (Génératives, Multiniveaux et Abstraites), en utilisant les transferts adéquats. Une gma-structure est plus « générale » et plus « générative » qu'une uma-structure, car ses niveaux de surface (fonctions syntaxiques, catégories syntagmatiques...) peuvent être vides, et sinon ne sont que des préférences indiquées par le transfert.

6. Pour chaque langue cible, la génération structurale produit à partir de la gma-structure cible une *uma-structure cible* homogène avec ce que serait le résultat de l'analyse (et de la désambiguïsation) du texte cible qui sera généré. Cette étape consiste à choisir la paraphrase à générer en calculant les niveaux de surface[27] et une première approximation de l'ordre des mots à partir des niveaux plus profonds (relations logiques et sémantiques, traits sémantiques, etc.).

---

25. Les deux premier modules peuvent travailler directement avec le texte contenu dans l'objet HyperCard. À partir du troisième, on travaille sur une transcription contenue dans un enregistrement du « fichier miroir » associé à la pile, ainsi que sur des résultats intermédiaires de traitement. Cela nous force à verrouiller le champ textuel original (sauf si l'auteur décide de le modifier et accepte de recommencer l'interaction depuis le début), et à éditer un « textoïde » image.

26. « Concrète » signifie que le texte original peut être retrouver à partir de la structure de façon directe (par un parcours préfixe des feuilles pour des structures de constituants et par un parcours infixe de tous les nœuds pour des structures de dépendances). Les nœuds et/ou les arcs de la structure peuvent contenir de l'information « de surface » aussi bien que « profonde » (organisation en prédicats/arguments, relations sémantiques...). Dans les structures « abstraites », les négations, les auxiliaires, les articles, etc., peuvent avoir été supprimés en tant que nœuds et être représentés dans les décorations, certains éléments élidés peuvent avoit été insérés, l'ordre peut avoit été normalisé, etc.

27. En particulier, les fonctions syntaxiques, les classes syntagmatiques, et les classes morphosyntaxiques, ces dernières en fonction des schémas dérivationnels des unités lexicales.

7. Le processus de traduction se termine par les générations syntaxiques et morphologiques. Quand tous les objets ont été traduits, on obtient la ou les piles images dans la ou les langues cibles.

8. Les uma-structures cibles peuvent être utilisées comme point de départ de *rétrotraductions* permettant à l'auteur (monolingue) de contrôler les traductions.

L'idée est que le système, traduisant par exemple du français au russe, retraduise au rédacteur ce qu'il va produire en russe, lui permettant ainsi de contrôler le résultat sans connaître un mot de russe. C'est d'ailleurs souvent ce qui se passe entre un interprète et son client.

Comme on suppose que la génération ne pose pas de problèmes, et qu'on veut éviter d'avoir à écrire tous les analyseurs dans une situation monosource et multicible, on ne repart pas du texte final en russe, mais de son uma-structure produite par la génération structurale. La rétrotraduction n'a pas à être identique à l'original pour que la traduction soit bonne, tout comme en traduction et surtout en interprétation humaine : on pourrait par exemple conserver actif et passif dans un sens (F-R), et les échanger dans l'autre (R-F) : il n'y aurait même pas de point fixe.

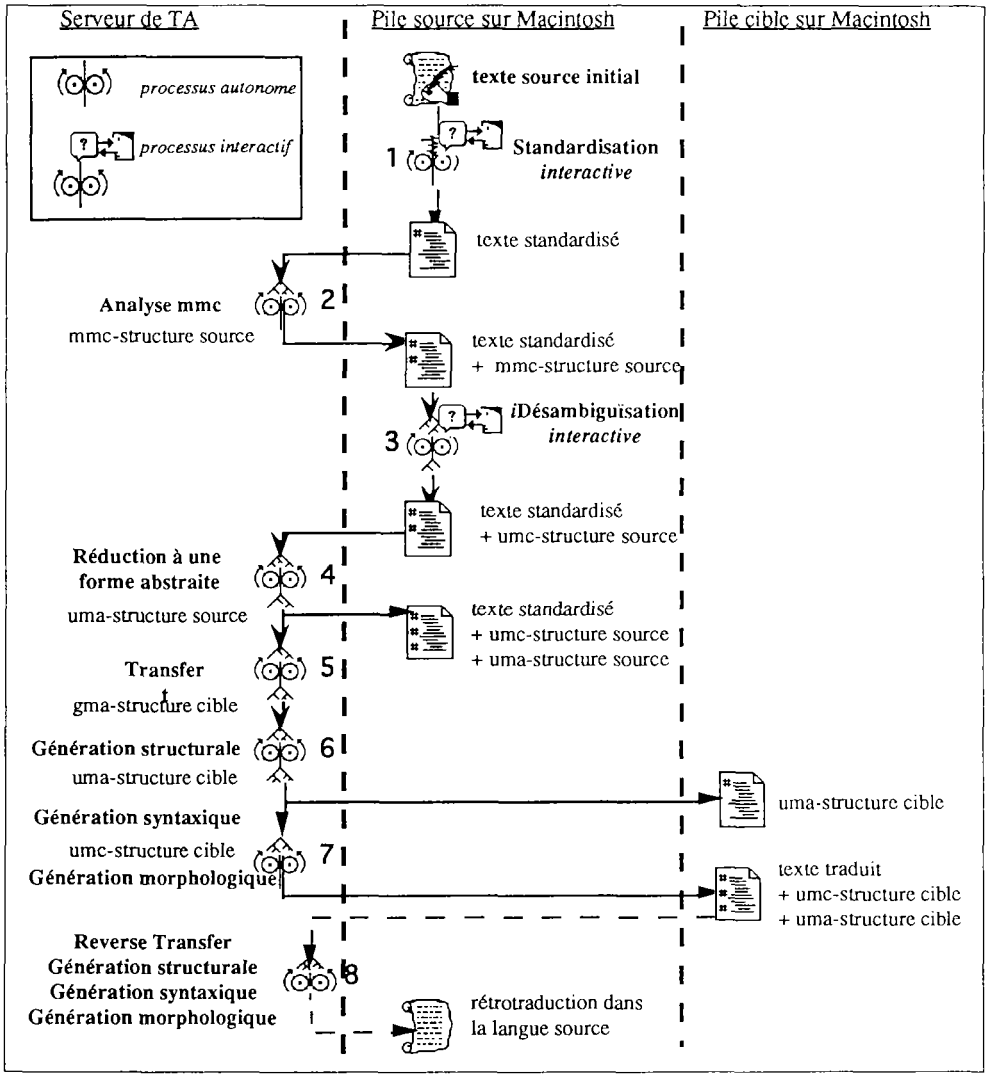


FIGURE 2 : Organisation générale du processus de traduction en LIDIA-1.

Implémentation

a. Répartition des tâches entre le serveur de TA et le Macintosh

Comme indiqué sur le diagramme, on trouve sur le serveur de TAO des « morceaux » de systèmes de TAFL (dictionnaires et grammaires écrits dans des langages de règles), écrits en Ariane-G5. Nous n'entrerons pas ici dans le détail.

Sur la station de rédaction, on trouve les modules mentionnés plus haut, ainsi que les ressources correspondantes (dictionnaire des syntagmes spéciaux, BDLM par acceptions interlingues) et les programmes propres à LIDIA. De plus, chaque pile à traduire est « doublée » *par un fichier miroir*, stockant les unités de traduction et les résultats des divers traitements.

Sur la station de rédaction comme sur le serveur de TAO, on trouve bien sûr des outils de gestion de la communication et de l'interaction.

b. Sur le Mac

HyperCard est le frontal de tout le système, mais, comme HyperTalk n'est pas assez puissant pour supporter certaines des tâches nécessaires, en particulier la génération des dialogues de désambiguïsation, nous avons écrit la plus grande partie de notre logiciel en CLOS (*Common Lisp Object System*). HyperCard et les programmes CLOS communiquent par des « AppleEvents », en utilisant le protocole standard IACP (*Inter Application Communication Protocol*) de Mac.OS-7. L'échange des données entre le Macintosh et le serveur de TA est implémenté à l'aide du kit de programmation d'Avatar (*Mac-MainFrame programmer's Toolkit*).

LIDIA-1 a 2 catégories principales d'objets, les *conteneurs* et les *contrôleurs*, avec 3 classes de conteneurs (`LIDIA-File`, `Mirror-Object`, `Disambiguation-Scheduler`) et 4 de contrôleurs (`LIDIA/HC-Communication-Controller`, `Remote-Translation-Jobs-Entry-Controller`, `Translation-Jobs-Fetch-Controller`, et `Translation-Process-Controller`). `LIDIA-File` a 5 instances, `Mirror-File`, `Translation-Jobs-in-Demand`, `To-Be-Fetched-Translation-Jobs`, `Prepared-Dialogues`, `Suspended-Treatments`. Ces fichiers contiennent toute l'information nécessaire sur l'état global de tout le système et sont constamment mis à jour (comme les piles HyperCard elles-mêmes).

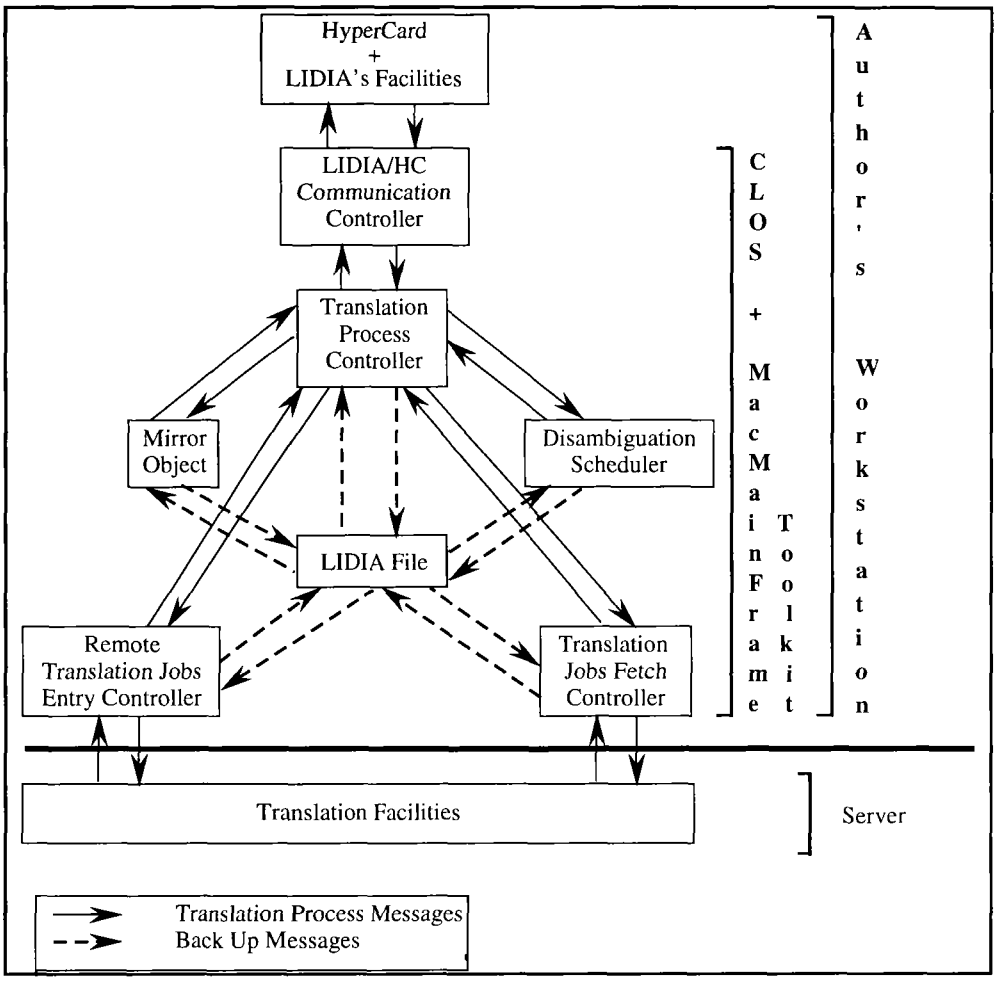


FIGURE 3 : Communication entre les objets de LIDIA-1.

Les *objets miroirs* contiennent toute l'information nécessaire au processus de traduction et à la construction des piles cibles.

Nous distinguons entre les *informations statiques* et les *informations dynamiques.*

Les informations statiques sont celles attachées par HyperCard à chaque objet. Les informations dynamiques sont celles utilisées par LIDIA pour traduire le contenu d'un objet.

Informations statiques :
- numéro de carte,
- localisation (fond ou carte),

- ID d'objet[28],
- type d'objet (champ ou bouton),
- fonte, taille et style,
- visibilité de l'objet.

Informations dynamiques :
- style d'énoncé/genre de texte,
- information textuelle sous la forme directement manipulée par HyperCard,
- transcription du texte, avec les annotations éventuelles,
- traitement à effectuer,
- langue(s) cible(s),
- étape courante du traitement,
- structures concrète et abstraite source,
- résultat(s) de traduction,
- uma-structure(s) cible(s) et rétrotraduction(s) éventuelle(s).

Interface utilisateur

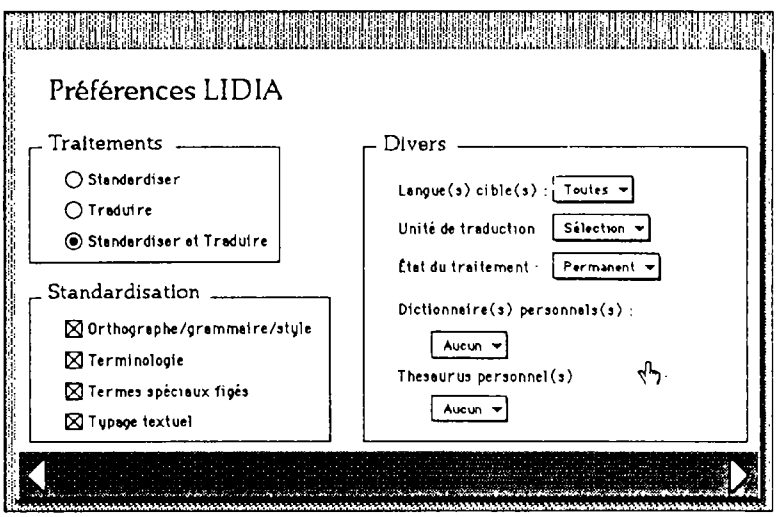


FIGURE 4 : Carte Préférences de LIDIA-1.

I. Préférences

Il y a quatre sortes de préférences relatives à la tâche, aux serveurs (de TA et de communication), à l'utilisation et aux ressources lexicales.

28. HyperCard affecte un numéro dit « ID d'objet » (*object ID number*) à chaque objet d'une pile. Ce numéro est unique pour chaque type d'objet à l'intérieur de l'objet qui le contient, ne change jamais et, si l'objet est supprimé, n'est pas réaffecté à un objet nouvellement créé. LIDIA utilise ces IDs d'objet de façon interne.

Les premières sont montrées dans la figure 4 : l'utilisateur choisit les langues cibles actives, l'unité de traduction (sélection, carte ou pile), et les traitements désirés : correction d'orthographe et de style, normalisation terminologique, syntagmes spéciaux figés, typage textuel et traduction.

Pour l'utilisateur, il s'agit du type de retour (sur demande ou automatique, un texte étant alors soumis dès que l'auteur quitte son conteneur), et du niveau de dialogue. Le profil lexical détermine le correcteur, les dictionnaires personnels, et les thesaurii à utiliser.

Si le module de catégorisation textuelle n'est pas utilisé, un type par défaut est attaché à tous les conteneurs de texte.

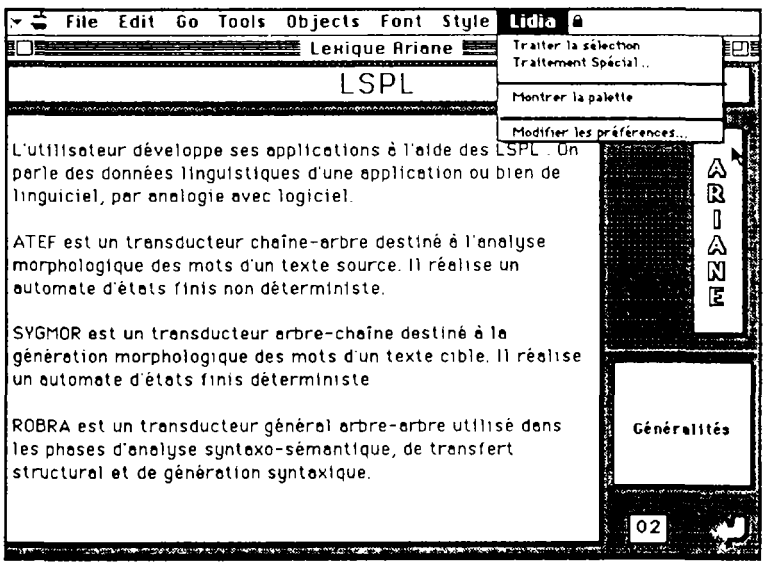


FIGURE 5 : Menu LIDIA.

II. Menu

LIDIA est accessible par le menu **Lidia**, qui change selon les préférences choisies.

Le menu montré ici offre quatre choix : traiter l'objet sélectionné selon les préférences en cours, ou le traiter avec d'autres préférences, montrer la palette, et modifier les préférences.

Quand l'auteur choisit l'un des deux premiers items, le curseur change de forme (√pour le premier cas, et √? pour le second) et on peut sélectionner l'objet à traiter.

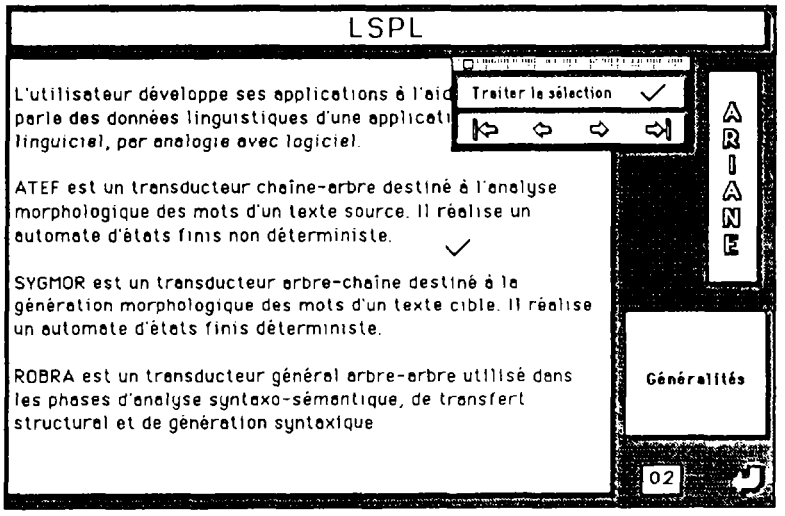


FIGURE 6 : Palette LIDIA-1.

III. Palette

Une palette est une fenêtre flottante qui est toujours au-dessus de toutes les fenêtres ouvertes d'une application.

En cliquant sur la partie du haut de la palette LIDIA, on active l'outil de traitement par défaut (√), et on sélectionne ensuite les objets à traiter.

En cliquant sur les icônes de la partie inférieure (⇦, ⇦, ⇨, ou ⇨), on va à la première carte de la pile, à la précédente, à la suivante, ou à la dernière.

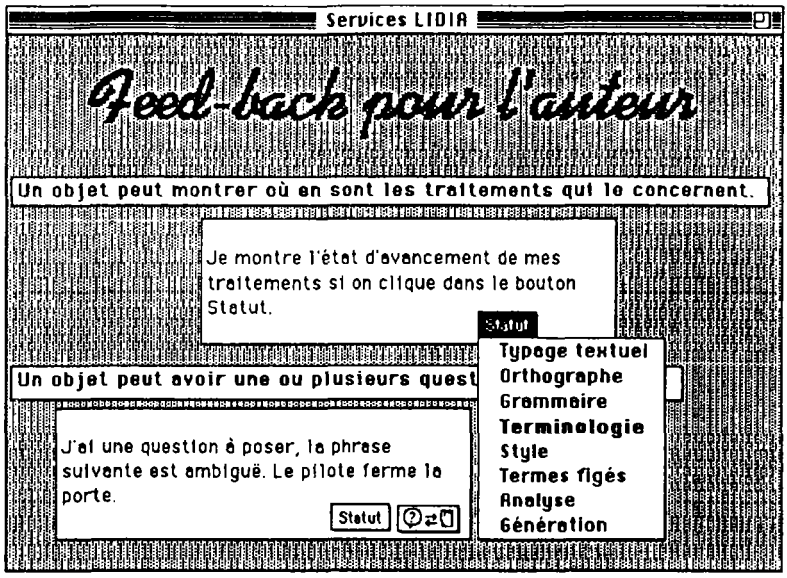


FIGURE 7 : Un observateur d'état de LIDIA-1.

IV. Contrôle

Un témoin d'état est un champ fugitif (*pop-up*) qui montre toutes les tâches, dans l'ordre d'exécution. La tâche en cours est affichée en gras et les suivantes en italique.

Quand des questions sur un objet sont prêtes à être posées, un bouton ( ) apparaît au-dessus de lui. Si l'auteur désire répondre, il clique sur ce bouton et le dialogue commence.

Quand la dernière réponse a été donnée, le traitement suivant peut commencer (si un dialogue est quitté sans être terminé, LIDIA-1 attend que l'auteur y revienne).

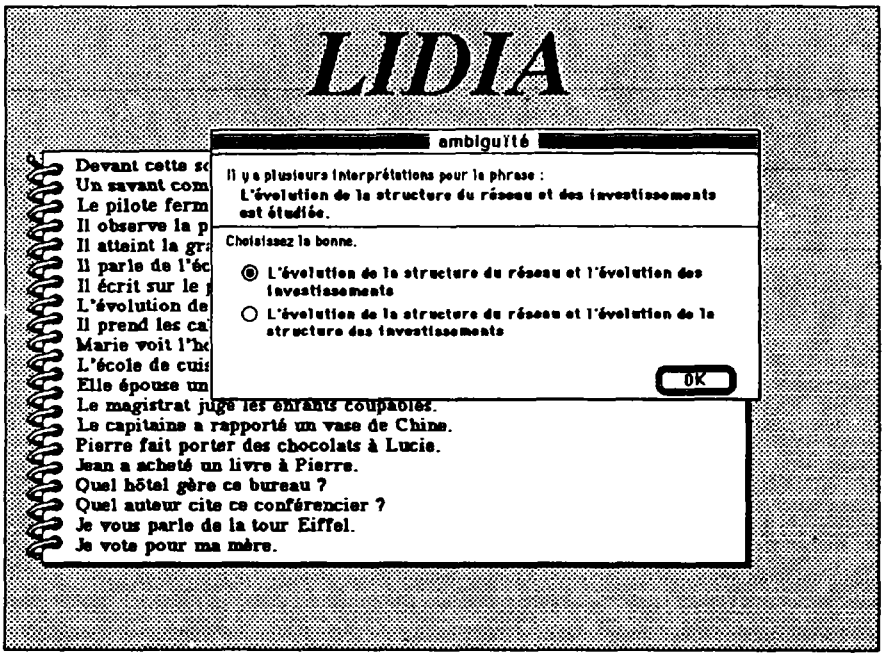


FIGURE 8 : Un dialogue de désambiguïsation.

V. Dialogues

Chaque question donne lieu à un dialogue par menu, illustré par l'image d'écran ci-dessus.

Dans cet exemple, on demande à l'auteur de choisir, grâce à deux paraphrases, l'interprétation correcte dans un cas d'ambiguïté de *géométrie* (voir *infra*).

Pour la clarification lexicale, on lui demande de choisir l'acception correcte d'un terme.

Pour la résolution d'anaphores, on lui demande de choisir le bon référent (en utilisant encore le paraphrasage).

c. *Encodage interne de textes multilingues dans un jeu de caractères universel*

Le premier problème est de traiter des textes dans des langues écrites avec des systèmes d'écriture variés. Jusqu'aux années 80, les systèmes informatiques n'offraient qu'un choix très réduit de jeux de caractères (comme romain sans diacritiques, mélange cyrillique/romain sans minuscules), de sorte qu'il était impératif d'utiliser des transcriptions. Dix ans après, presque tous les constructeurs d'ordinateurs commencèrent à offrir des jeux de caractères étendus et les systèmes d'exploitation *localisés.*

Mac.OS-7.1, fondé sur Unicode (un sous-ensemble de la norme multiscript ISO-10646) et disponible depuis fin 1992, est le premier système d'exploitation réellement multiscript[29] : avec n'importe quel texteur utilisant le *Script Manager* standard, il est possible de composer un document contenant des parties dans presque toutes les langues d'Europe, ainsi qu'en arabe, japonais, chinois, etc.

Cependant, Mac.OS-7.1 est encore un cas unique[30], et le problème reste entier si l'on veut échanger du texte entre ordinateurs de différentes marques, ou simplement transmettre du texte via les réseaux télématiques. Par exemple, l'ASCII français n'est pas le même sur un PC et sur un Macintosh. Dans notre cas, le serveur de TA n'utilise même pas de l'ASCII, mais de l'EBCDIC.

Notre solution est d'utiliser des transcriptions romaines pour les représentations internes des textes, des grammaires et des dictionnaires. Une transcription consiste en un jeu de caractères et une méthode pour représenter le matériau textuel considéré (pas seulement les mots, mais aussi la structure logique, la mise en page, et éventuellement d'autres informations), en n'utilisant que ces caractères. Le jeu de caractères et la méthode à utiliser dépendent de l'importance relative accordée à la portabilité, à la lisibilité et à la compacité.

Pour la TA, nous avons depuis longtemps utilisé un jeu de caractères de transcription presque identique à celui de PL/I (ni minuscules ni diacritiques, mais seulement les majuscules simples, les signes de ponctuation usuels et quelques signes spéciaux). Cela donne une portabilité totale[31], aux dépens de la lisibilité.

Par exemple, `"*A!2 *NOE!4L , *MAC*ALLISTER VA AUX **USA"`
code : `"À Noël, MacAllister va aux USA".`

Pour la partie micro de LIDIA-1, la lisibilité est plus importante, et nous utilisons un sous-ensemble plus grand de l'ISO-646, contenant les majuscules et les minuscules, mais pas les diacritiques. Les diacritiques sont représentés par des *séquences spéciales* introduites par « ! » parce que, dans la documentation technique, les pièces sont souvent référencées par des identificateurs où lettres et chiffres sont mélangés (comme : « XA1 »). L'information relative à la structure ou à la mise en page du texte est représentée par des marques, ou *balises,* dans l'esprit de SGML et de la TEI (`<parag>`, `<section>`, `<greek>`, etc.). On indique de façon analogue un changement de langue et/ou de système d'écriture (certaines langues en utilisent plus d'un).

d. Encodage d'informations linguistiques pouvant provenir d'une préédition (indirecte)

Les syntagmes figés spéciaux sont transformés en des occurrences spéciales, de façon à aider l'analyse et la traduction. Par exemple, « Cacher les bulles » devient

29. Mac.OS-7.1 n'est pas encore lui-même multilingue : bien que tous ses utilitaires soient indépendants des langues, chaque version distribuée n'a les messages et autres ressources spécifiques des langues que dans une langue.
30. Le « documenteur » Star™ de Xerox a été le seul outil vraiment multilingue jusqu'à la sortie de WinText™ sur Macintosh en 1987, mais les systèmes d'exploitation sous-jacents étaient respectivement strictement monolingues, ou seulement localisables.
31. Dans beaucoup de pays, comme la Thaïlande, les minuscules romaines sont remplacées par les caractères locaux dans les terminaux bilingues.

`&FXN_Cacher_les_bulles`, qui peut être traité par une sous-grammaire morphologique appropriée.

Après la désambiguïsation lexicale, nous pourrions attacher à chaque occurrence le numéro du sens dans la BDLM, par exemple `glace.1` pour *glace à manger* et `glace.2` pour *miroir*. Mais cela n'est pas très lisible et interdit la préédition directe. Nous permettons donc d'ajouter au numéro de sens un fragment de la définition qui donne la distinction, la *clé sémantique,* d'habitude un autre mot ou terme, ce qui donne `glace.1=aliment` ou `glace.2=miroir`.

Dans le futur, LIDIA devrait permettre des présentations alternatives plus lisibles, utilisant, par exemple, les majuscules (et « * » pour indiquer les *vraies* majuscules) pour le texte lui-même, et les minuscules pour les annotations (par exemple, `GLACE.1=aliment`). Comme en général plus d'une clé sémantique peut être associée à une même acception, par exemple `glace.1=a_manger` ou `glace.1=dessert`, il faudra aussi permettre à l'utilisateur d'entrer `glace=dessert`, et que le système consulte la BDLM, trouve que l'acception de *glace* la plus proche de *dessert* est le sens numéro 1, et transforme `glace=dessert` en `glace.1=dessert` ou même en `glace.1=aliment`.

Les annotations concernant les informations grammaticales relatives aux mots, comme la catégorie morphosyntaxique (verbe, nom...), le nombre, le genre, le temps, le mode, etc., sont attachées aux occurrences de façon analogue.

Le dernier type d'annotations concerne les structures concrètes (mmc ou umc). Pour délimiter les groupes syntagmatiques, nous utilisons des parenthèses spéciales, comme `{&rel...}` pour une proposition relative[32], ou simplement `{...}` si la catégorie syntagmatique n'est pas connue ou trop ésotérique pour les utilisateurs naïfs (par exemple, *groupe adjectival* ou *groupe cardinal*). Pour représenter un lien anaphorique, on attache au pronom une copie de son référent. En cas d'élision, on rajoute des occurrences *cachées* (`centrale .&eld=inertielle`). D'autres informations grammaticales et sémantiques peuvent être attachées aux terminaux et aux non-terminaux.

Par exemple, « Devant cette somme, il ne rend pas sa glace immangeable pour autant » pourrait avoir la représentation intermédiaire suivante (avant d'avoir fini la désambiguïsation) :

```
{{&grd,cause *devant.&vrb cette somme.2&nf , } il ne rend.2
pas=ne sa glace.1 immangeable pour autant }
```

qui lèverait l'ambiguïté sur *devant* (verbe/préposition) et *rendre* (vomir/restituer/faire devenir), d'où en anglais :

Owing this sum, he doesn't vomit his unedible ice cream for that reason, ou
Owing this amount, he doesn't vomit his unedible ice cream for that reason,

plutôt que, entre autres,

Owing this amount, he doesn't render his ice cream unedible for that reason, ou
Facing this summa [Opus Magnum], he doesn't give back his mirror for that reason.

32. Ou bien on ajoute « rel » à l'information attachée à sa tête, comme dans l'exemple suivant (« compose.&v,phvb »).

Le point principal est que *le système d'annotations concerne plusieurs niveaux de description linguistique, mais est incomplet à chaque niveau,* parce qu'aucune notion non familière ne doit apparaître. Par exemple, *verbe* est une notion familière pour presque tout adulte instruit, mais pas *verbe modal.* Au niveau des fonctions syntaxiques, *sujet, objet* et *complément* sont familiers, mais sans doute pas *attribut, épithète, tête* (ou *gouverneur*). Il en va de même au niveau des cas profonds.

Après désambiguïsation, on obtient un nouveau texte annoté par projection systématique de la umc-structure (qui doit contenir des informations plus complètes et détaillées). On le garde à part, pour que les utilisateurs expérimentés puissent l'éditer directement. Voici à titre d'exemple un texte de 3 phrases contenu dans la pile de documentation d'Ariane-G5.

```
Un processus de traduction en ARIANE-G5 se compose d'une suite de trois
étapes (analyse, transfert et génération). Chaque étape est constituée
d'une suite de différentes phases de traitement. Chaque phase est relative
à l'emploi d'un LSPL précis.
```

Et voici les présentations actuelle et future de la forme annotée :

```
{ { *un.&art processus.&n,suj { de.&prep traduction.&n,comp { en.&prep
**ariane-g5.&np,comp } } } se.&refl compose.&v,phvb { d'une.&art
suite.&n,obj1 { de.&prep trois.&card e!1tapes.&n,comp { (.&lp
analyse.&n,app ,.&ponc transfert.&n,coord et.&cjcoord
ge!1ne!1ration.&n,coord ).&rp } } } ..&ponc } { { *chaque.&art
é!1tape.&n,suj } est.&v,aux constitue!1e.&v,phvb { d'.&prep une.&art
suite.&n,comp { de.&prep { diffe!1rentes.&adj,epit } phases.&n,comp {
de.&prep traitement.&n,comp } } } ..&ponc } { { *chaque.&art phase.&n,suj }
est.&v,phvb { relative.&adj,atsubj { a!2.&prep l'.&art emploi.&n,obj1 {
d'.&prep un.&art **lspl.&np,comp { pre!1cis.&adj,epit } } } } ..&ponc }
```

```
{ { *UN.&art PROCESSUS.&n,suj { DE.&prep TRADUCTION.&n,comp { EN.&prep
**ARIANE-G5.&np,comp } } } SE.&refl COMPOSE.&v,phvb { D'UNE.&art
SUITE.&n,obj1 { DE.&prep TROIS.&card E!1TAPES.&n,comp { (.&lp
ANALYSE.&n,app ,.&ponc TRANSFERT.&n,coord ET.&cjcoord
GE!1NE!1RATION.&n,coord ).&rp } } } ..&ponc } { { *CHAQUE.&art
E!1TAPE.&n,suj } EST.&v,aux CONSTITUE!1E.&v,phvb { D'.&prep UNE.&art
SUITE.&n,comp { DE.&prep { DIFFE!1RENTES.&adj,epit } PHASES.&n,comp {
DE.&prep TRAITEMENT.&n,comp } } } ..&ponc } { { *CHAQUE.&art PHASE.&n,suj }
EST.&v,phvb { RELATIVE.&adj,atsubj { A!2.&prep L'.&art EMPLOI.&n,obj1 {
D'.&prep UN.&art **LSPL.&np,comp { PRE!1CIS.&adj,epit } } } } ..&ponc }
```

Le codage interne est bien plus qu'une question technique de second ordre, comme on le pense souvent. Sa définition n'est pas seulement très importante pour les développeurs de grammaires et de dictionnaires, mais, pour la concevoir de façon cohérente, il faut comprendre le fonctionnement interne d'un système de TA, et, dans notre cas, satisfaire la *contrainte d'accessibilité* (par des utilisateurs naïfs). C'est aussi un défi pour des linguistes habitués à faire des distinctions très subtiles que de devoir bâtir des systèmes n'utilisant que des informations *rustiques* obtenables de non-spécialistes. Or, cela est nécessaire pour que la TAFD *pour tous* puisse réussir.

Enfin, notons que cette idée (de préédition indirecte et/ou directe) est aussi utilisée dans d'autres projets, comme le projet LMT d'IBM (Rimon, McCord, Schwall *et al.* 1991)[33].

Dialogues de désambiguïsation

Dans la maquette LIDIA-1, les dialogues sont conduits uniquement à l'écran. Nous espérons expérimenter dans le futur l'introduction d'autres média, et notamment de synthèse vocale.

Pour ne pas surcharger les utilisateurs avec des choses nouvelles à apprendre, nous avons préféré nous en tenir à des dialogues par menu. Par exemple, nous avions pensé proposer un outil de manipulation graphique des structures concrètes, mais des utilisateurs potentiels et des ergonomes ont trouvé cela plus difficile que de choisir entre des paraphrases textuelles avec mise en relief des différences.

Enfin, une suggestion (Chandler *et al.* 1987) était de retarder toutes les interactions jusqu'au transfert. Compte tenu de nos objectifs de grande couverture et de *rusticité,* il nous a paru préférable de résoudre dès que possible les ambiguïtés impossibles ou très difficiles à résoudre automatiquement par l'un des processus ultérieurs.

Classification des ambiguïtés

Rappelons que les *ambiguïtés lexicales* concernent non seulement la classique polysémie de termes (par exemple, *diplôme* se traduit par *diploma* ou *degree*), mais aussi les *ellipses lexicales*[34]. Dans les deux cas, LIDIA construit un menu avec les choix possibles, rangés dans l'ordre de leurs poids courants.

Les autres ambiguïtés présentes dans la mmc-structure sont partitionnées en trois classes :

---

33. « The user can mark the input string selectively with brackets <...> (to any degree) to force parsing choice and desambiguate the input. "User" in this context can also apply tools (such as the interactive disambiguator) which may introduce such marks in their output. »

34. Supposons qu'un texte parle d'un vaisseau spatial contenant une « centrale électrique » (*electric plant*) et une « centrale inertielle » (*inertial guidance system*). La forme complète est souvent remplacée par la forme élidée (« centrale »). Bien qu'il soit crucial de désambiguïser pour traduire correctement (par les formes élidées correspondantes, « *plant* » ou « *system* »), on ne connaît aucune solution automatique. Une occurrence donnée de « centrale » peut être ou ne pas être une élision. Et si elle l'est, il est encore plus difficile de rechercher un candidat pour la forme complète dans un hypertexte que dans un texte usuel.

- Il y a *ambiguïté de classe syntaxique* si deux classes syntaxiques ou plus sont affectées à une occurrence dans l'ensemble des solutions produites par l'analyseur.
- Il y a *ambiguïté de géométrie* si les structures arborescentes de deux solutions donnant les mêmes valeurs de classes aux occurrences sont différentes.
- Il y a *ambiguïté de fonction syntaxique et/ou de relation sémantique* si l'analyseur produit deux structures de même classe et de même géométrie, différant donc par l'information portée par certains nœuds non terminaux.

Si plusieurs problèmes apparaissent dans la même phrase (énoncé), nous utilisons la stratégie suivante :

1. déterminer la segmentation correcte en groupes simples ;
2. déterminer les arguments et les circonstants de chaque prédicat commun à deux solutions ;
3. déterminer les relations syntaxiques et sémantiques correctes entre les groupes simples.

C'est l'ordre naturel que suivent les humains en cas de problème, et il est important que les utilisateurs comprennent facilement et clairement ce que fait le système. Au niveau du système, cela revient à résoudre dans l'ordre les ambiguïtés de classe, de géométrie, et de relations.

Production des dialogues

a. Raffinement des classes et schémas de problèmes
Nous ne pouvons pas proposer une méthode de désambiguïsation unique pour chaque classe d'ambiguïté, en particulier à cause des croisements multiples. D'autre part, nous nous sommes donné la contrainte de produire des menus proposant des choix entre des paraphrases. En examinant un grand nombre de configurations d'ambiguïtés, nous sommes arrivés à distinguer huit types de problèmes, pour lesquels on sait produire des paraphrases *désambiguïsantes* :

A. ambiguïté de classe syntaxique :
1. ambiguïté de classe syntaxique sans groupes coordonnés ambigus

Le pilote ferme la porte :	*The firm pilot carries her.*
	The pilot shuts the door.

2. ambiguïté de classe syntaxique associée à un groupe coordonné

Il regarde la photo et la classe :	*He looks at the photograph and the class.*
	He looks at the photograph and files it.

B. ambiguïté de géométrie :
3. ambiguïté de structure argumentaire du verbe

Il parle de l'école de cuisine :	*He talks about the cooking school.*
	He talks from the cooking school.
	He talks from the school about cooking.

4. ambiguïté de coordination

Il prend des crayons et des cahiers noirs :	*He takes pencils and black notebooks.*
	He takes black pencils and black notebooks.

5. ambiguïté de subordination

L'école de cuisine lyonnaise est fermée :	*The lyonnaise cooking school is closed.*
	The school of lyonnaise cooking is closed.

C. ambiguïté de fonction syntaxique et/ou de relation logico-sémantique

6. ambiguïté de relation logico-sémantique

Pierre fait porter des chocolats à Lucie :	*Pierre is having chocolates sent to Lucie.*
	Pierre is having chocolates sent with Lucie.

7. ambiguïté sur les positions d'arguments d'un verbe transitif direct

Quel auteur cite ce conférencier :	*Which author is this lecturer quoting?*
	Which lecturer is this author quoting?

8. ambiguïté de fonction syntaxique

Il parle de la tour Eiffel :	*He is talking about the Eiffel Tower.*
	He is talking from the Eiffel Tower.

b. Description des opérateurs de base

Pour produire ces paraphrases, nous utilisons les six opérateurs de base suivants, qui opèrent sur des fragments de la mmc-structure :

`Generate` produit une forme fléchie à partir d'une unité lexicale.

`Distribute` distribue une occurrence ou un groupe d'occurrences sur d'autres groupes d'occurrences, et établit un lien entre les résultats des distributions locales.

exemple : `Distribute(A, B C, D, 1, 2, ou, 1, 3) -> A B C ou A D`

`Permute` change l'ordre d'une liste d'occurrences.

exemple : `Permute( A, B, C D, E, F, G, 1, 2, 5, 6) -> A B F G`

`Substitute` substitue à un terme un autre terme recherché dans la BDLM en fonction de conditions.

`Project` appliqué à un nœud de l'arbre, projette la classe syntaxique, et, en fonction de cette classe, certaines informations contenues dans la BDLM.

`Bracket` met entre parenthèses la sous-liste d'une liste d'occurrences comprise entre deux bornes.

c. Discussion

De nombreux projets, en particulier (Kay 1973 ; Melby 1981 ; Chandler *et al.* 1987 ; Sadler 1989 ; Brown et Nirenburg 1990 ; Huang 1990 ; Maruyama *et al.* 1990 ; Wehrli 1990), ont expérimenté diverses stratégies de désambiguïsation. Si aucun système utilisable en pratique n'en est sorti, c'est à notre avis surtout parce que le processus de désambiguïsation apparaît comme une boîte noire d'où sortent des questions assez imprévisibles et ésotériques. Nous avons donc consciemment préféré la rusticité à la sophistication, et cherché à construire un système de structure simple, posant des questions compréhensibles par tout bachelier.

Pour l'instant, nous utilisons une stratégie de désambiguïsation *câblée* dans les programmes de LIDIA-1. Or, il est impossible d'affirmer que c'est *la meilleure* (et nous doutons fort qu'il en existe une). En tout état de cause, les utilisateurs devraient pouvoir choisir entre divers modes de désambiguïsation. C'est pourquoi nous travaillons sur un outil qui permettrait à des linguistes et des ergonomes de définir et d'expérimenter des stratégies de désambiguïsation variées, dans divers contextes (type d'utilisateur, autre média...).

Une première étape a été la construction d'un ensemble d'opérateurs de base pour la génération de paraphrases. La suivante devrait permettre aux linguistes de définir eux-mêmes les schémas de problèmes. Un but plus lointain serait d'offrir des outils de *programmation ambiguë*, permettant de décrire les types d'ambiguïtés et les stratégies de désambiguïsation (y compris le recours à l'auteur ou à une ontologie) à l'intérieur des grammaires du système de TA lui-même.

Conclusion

Le concept de TAFD cristallise des idées venant de systèmes et de recherches antérieurs (critique textuelle, TAFL interactive, TAFL avec préédition, TAFC avec augmentation, langages contrôlés, sous-langages...). Cependant, la contrainte d'interagir avec un auteur n'ayant aucune connaissance des langues cibles ni de la linguistique permet de parler d'un nouveau paradigme.

Comme nous l'avons dit, les paradigmes ne sont pas exclusifs, et ont chacun leurs domaines d'emploi, qui peuvent se recouvrir plus ou moins. Il s'agit en effet de paradigmes techniques, et non de théories scientifiques. Mais la TAO est une technologie scientifique, qui bénéficie de temps en temps de concepts innovateurs provenant des sciences qui la sous-tendent, en en tirant des progrès incrémentaux, et qui peut aussi proposer aux théoriciens des problèmes intéressants.

En TAFD, les progrès incrémentaux dont nous parlons pourraient venir de la mise en œuvre d'approches intermédiaires nouvelles, comme :

- *le transfert multiniveau par acceptions interlingues ;*
- l'approche par *langage guidé* (préférences lexicales, styles d'énoncés, genres de textes), avec la combinaison de techniques symboliques et numériques ;
- *l'accessibilité et la rusticité des connaissances linguistiques*, permettant à terme l'enrichissement par l'utilisateur.

Parmi les problèmes nouveaux posés par la TAFD, les plus importants (et difficiles) nous semblent être :

- *le typage automatique de textes et d'énoncés,* ainsi que *le guidage de l'auteur* pour la partie interactive ;
- la conception (et la validation) de techniques de *désambiguïsation interactive multimédia ;*
- *l'enrichissement* du système par l'utilisateur, combiné avec un *réglage automatique.*

Il ne faut bien sûr pas confondre une étape d'étude et de maquettage avec la réalisation d'un produit. D'abord, il n'est pas exclu que la recherche bute sur des obstacles non prévus ou sous-estimés. Cela est arrivé plus d'une fois dans l'histoire de la TAO. Ensuite, dans le cas de la TAFD *pour tous*, le facteur d'échelle sera très grand (de l'ordre de 1 000 pour les bases de données lexicales) : même si la plupart des problèmes étaient résolus au niveau de LIDIA-1, rien ne dit que les méthodes de développement de bases de données linguistiques auront assez progressé pour qu'un système opérationnel de TAFD destiné au grand public soit techniquement réalisable et économiquement viable.

Cependant, l'avancement actuel des connaissances et des techniques nous permet d'être raisonnablement optimistes. Les enjeux scientifiques sont importants, et les enjeux économiques et culturels encore plus. En effet, si elle est faisable, la TAFD permettra de résoudre, au moins en partie, le problème crucial rencontré par les promoteurs des langues nationales en général, et par les défenseurs de la francophonie en particulier, à savoir l'impossibilité absolue, pour le plus grand nombre, d'écrire dans sa langue et d'être traduit dans d'autres dans un délai raisonnable et avec une garantie suffisante de qualité.

Remerciements

Je tiens à remercier ici H. Blanchon, qui a réalisé les figures et images d'écran de la quatrième partie, ainsi que la programmation non linguistique de LIDIA-1 sur le Macintosh. Merci également à tous les collègues du GETA qui ont participé à ce maquettage : M. Axtmeyer, E. Blanc, N. Denos, J.-Ph. Guilbaud, P. Guillaume, M. Lafourcade, D. Levenbach, N. Nédobejkine, F. Peccoud, B. Roudaud, M. Quézel-Ambrunaz, G. Sérasset, ainsi qu'à WinSoft et à Machina Sapiens, qui nous ont permis d'utiliser leurs logiciels gracieusement. Merci enfin (*last, but not least*), à A. Clas et P. Bouillon, pour m'avoir aimablement invité à présenter le paradigme de la TA fondée sur le dialogue dans l'ouvrage qu'ils ont conçu et par suite à ces journées.

Références

ABBOU, A. (dir.) (1988) : *Traduction Assistée par Ordinateur, Actes du séminaire international sur la TAO et dossiers complémentaires,* Observatoire des Industries de la Langue (OFIL), Paris, mars 1988, 234 p.

BLANC, E. et C. BOITET (dir.) (1990) : *DBMT-90, Post-COLING Seminar on Dialogue-Based MT (Machine Translation off/with dialogues). Organization, General Spirit, Final Program & Content of the 8 Sessions*, Le Sappey, 26-28 août 1990, 328 p.

BLANCHON, H. (1990) : « LIDIA-1 : Un prototype de TAO personnelle pour rédacteur unilingue », *Actes des X^e Journées sur les systèmes experts et leurs applications. Conférence spécialisée : Le traitement automatique des langues naturelles et ses applications,* Avignon, 28 mai-1 juin 1990, EC2, pp. 51-60.

BLANCHON, H. (1992) : « A Solution for the Problem of Interactive Disambiguation », *Proceedings of the 14th International Conference on Computational Linguistics,* COLING-92, Nantes, pp. 1233-1238.

BLANCHON, H., GUILBAUD, J. P. et N. NÉDOBEJKINE (1992) : « LIDIA: the Disambiguation Process – le processus de désambiguïsation », Exposition COLING-92, Nantes, 23-28 juillet 1992, Pile HyperCard.

BOITET, C. (1985) : « Traduction (assistée) par ordinateur : ingénierie logicielle et linguicielle », *Proceedings Colloque RF&IA*, Grenoble, AFCET.

BOITET, C. (1988a) : *Hybrid Pivots using M-structures for multilingual Transfer-Based MT Systems*, Jap. Inst. of Electr., Inf. & Comm. Eng., June 1988, NLC88-3, pp. 17-22.

BOITET, C. (1988b) : « PROs and CONs of the Pivot and Transfer Approaches in Multilingual Machine Translation », *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 18-19 August 1988, BSO, 13 p.

BOITET, C. (1988c) : « Representation and Computation of Units of Translation for Machine Interpretation of Spoken Texts », *Computer & Artificial Intelligence*, 6, pp. 505-546 (and TR-I-0035, ATR, Osaka).

BOITET, C. (1988d) : « Software and Lingware Engineering in Modern M(A)T Systems », *in* Bátori (dir), *Handbook for Machine Translation*, Niemeyer.

BOITET, C. (1989a) : « Motivation and Architecture of the LIDIA Project », *Proceedings MTS-II (MT Summit)*, Munich, 16-18 août 1989, 12 p.

BOITET, C. (1989b) : « Speech Synthesis and Dialogue-Based Machine Translation », *Proceedings of the ATR Symposium on Basic Research for Telephone Interpretation*, Kyoto, December 1989, 12 p.

BOITET, C. (1990a) : « Multilingual Machine Translation Does not Have to be Saved by Interlingua », *Proceedings of MMT'90*, Tokyo, 5-6 Nov. 1990, 2 p.

BOITET, C. (1990b) : « Towards Personal MT: On Some Aspects of the LIDIA Project », H. Karlgren (dir.), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, 20-25 août 1990, ACL, vol. 3/3, pp. 30-35.

BOITET, C. (1992) : « TAO personnelle et promotion des langues nationales », *Turjumãn*, revue de traduction et d'interprétation, École Supérieure Roi Fahd de Traduction, Université Abdelmalek Essaâdi, Tanger, 1/1, avril 1992, pp. 35-49.

BOITET, C. et Y. ZAHARIN (1988) : « Representation Trees and String-tree Correspondences », D. Várgha (dir), *Proceedings of the 12th International Conference on Computational Linguistics*, COLING-88, Budapest, 22-27 Aug. 1988, ACL, pp. 59-64.

BOURBEAU, L. (1990) : « Élaboration et mise au point d'une méthodologie d'évaluation linguistique de systèmes de traduction assistée par ordinateur », *Rapport final*, Secrétariat d'État du Canada, Langues officielles et traduction, Direction de la planification, gestion et technologie, Québec, mars 1990, 203 p.

BROWN, R. D. (1989) : « Augmentation », *Machine Translation*, 4, pp. 1299-1347.

BROWN, R. D. et S. NIRENBURG (1990) : « Human-Computer Interaction for Semantic Disambiguation », H. Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, ACL, vol. 3/3, pp. 42-47.

CARBONNELL, J. G. et M. TOMITA (1985) : « New Approaches to Machine Translation », S. Nirenburg (dir.), *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-85, Hamilton, N.Y., 14-16 Aug. 1985, pp. 59-74.

CHANDIOUX, J. (1988) : « 10 ans de METEO (MD) », A. Abbou (dir.), *Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires*, Observatoire des Industries de la Langue (OFIL), Paris, mars 1988, pp. 169-173.

CHANDIOUX J. et M.-F. GUÉRARD (1981) : « METEO : un système à l'épreuve du temps », *Meta*, 26-1, pp. 17-22.

CHANDLER, B., HOLDEN, N., HORSFALL, H., POLLARD, E. et M. MCGEE WOOD (1987) : *N-tran Final Report*, Alvey Project, 87/9, CCL/UMIST, Manchester.

COUTAZ, J. (1988) : *Interface Homme-ordinateur : Conception et Réalisation*, Thèse d'État, Université Joseph Fourier, Grenoble.

COWIE, J., GUTHRIE, J. et L. GUTHRIE (1992) : « Lexical Disambiguation Using Simulated Annealing », C. Boitet (dir), *Proceedings of the 14th International Conference on Computational Linguistics*, COLING-92, Nantes, vol. 1/4, pp. 359-365.

DESCLÉS, J.-P. (1987) : « Sémantique », *Technologos*, LISH-CNRS, printemps 1987.

DUCROT, J.-M. (1982) : « TITUS IV », P. J. Taylor (dir), *Information Research in Europe. Proceedings of the EURIM 5 Conference (Versailles)*, ASLIB, London.

DUCROT, J.-M. (1988) : « Le système TITUS IV », A. Abbou (dir), *Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires*, Observatoire des Industries de la Langue (OFIL), Paris, mars 1988, pp. 55-71.

GERBER, R. et C. BOITET (1985) : « On the Design of Expert Systems Grafted on MT Systems », S. Nirenburg (dir), *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-85, Hamilton, N.Y., 14-16 Aug. 1985, Colgate University, pp. 116-134.

HEIDORN, G. E., JENSEN, K. et L. A. MILLER (1982) : « The EPISTLE Text-Critiquing System », *IBM System Journal*, 21/1, pp. 305-326.

HIRAKAWA, H., NOGAMI, H. et S.-Y. AMANO (1991) : « EJ/JE Machine Translation System AS-TRANSAC - Extension toward Personalization », *Proceedings MTS-III* (MT Summit), Boston, 1-4 July 1991, vol. 1/1, pp. 73-80.

HUANG, X. M. (1990) : « A Machine Translation System for the Target Language Inexpert », H. Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, 20-25 Aug. 1990, ACL, vol. 3/3, pp. 364-367.

HUTCHINS, W. J. (1986) : *Machine Translation : Past, Present, Future*, Chichester, England, Ellis Horwood Limited, 382 p.

HUTCHINS, W. J. et H. L. SOMERS (1992) : *An Introduction to Machine Translation*, Londres, Academic Press.

ISABELLE, P. et L. BOURBEAU (1984) : « TAUM-AVIATION : its Technical Features and Some Experimental Results », *Computational Linguistics*, 11/1, pp. 18-27.

JEIDA (1989) : *A Japanese View of Machine Translation in Light of the Considerations and Recommendations Reported by ALPAC, USA*, Japanese Electronic Industry Development Association, Tokyo, 197 p.

KAY, M. (1973) : « The MIND System » , R. Rustin (dir), *Current Computer Science Symposium 8: Natural Language Processing*, New York, Algorithmics Press, Inc., pp. 155-188.

KAY, M. (1980) : « The Proper Place of Men and Machines in Language Translation », *Research Report*, CSL-80-11, Xerox, Palo Alto Research Center, Oct. 1980.

KITTREDGE, R. (1983) : « Sublanguage – Specific Computer Aids to Translation – a survey of the most promising application areas », *Contract*, n° 2-5273, Université de Montréal et Bureau des Traductions, mars 1983, 95 p.

KITTREDGE, R. (1986) : « Analyzing Language in Restricted Domains », R. Grishman et R. Kittredge (dir), *Sublanguage Description and Processing*, Hillsdale, Lawrence Erlbaum, New Jersey.

LEHRBERGER, J. et BOURBEAU L. (1988) : *Machine Translation. Linguistic Characteristics of MT Systems and General Methodology of Evaluation*, John Benjamins, 240 p.

MARUYAMA, H., WATANABE, H. et S. OGINO (1990) : « An Interactive Japanese Parser for Machine Translation », H. Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, 20-25 Aug. 1990, ACL, vol. 2/3, pp. 257-262.

MELBY, A. K. (1981) : « Translators and Machines - Can they Cooperate? », *Meta*, 26-1, pp. 23-34.

MELBY, A. K. (1982) : « Multi-Level Translation Aids in a Distributed System », *Proceedings of the International Conference on Computational Linguistics*, COLING-82, Prague, 5-10 juillet 1982, vol. 1/2, pp. 215-220.

MELBY, A. K., SMITH, M. R. et J. PETERSON (1980) : « ITS : An Interactive Translation System », M. Nagao (dir), *Proceedings of the International Conference on Computational Linguistics*, COLING-80, Tokyo, 30 septembre-4 octobre 1980, pp. 424-429.

NIRENBURG, S. (1989) : « Knowledge-Based Machine Translation », *Machine Translation*, 4, pp. 5-24.

NIRENBURG, S. *et al.* (1989) : *KBMT-89 Project Report*, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, April 1989, 286 p.

NIRENBURG, S. et K. GOODMAN (1990) : « Treatment of Meaning in MT Systems », *Proceedings ROCLing-III*, Taipeh, 20-22 Aug. 1990, pp. 83-101.

NYBERG, E. H. et T. MITAMURA (1992) : « The KANT System : Fast, Accurate, High-Quality Translation in Practical Domains », C. Boitet (dir), *Proceedings of the 14th International Conference on Computational Linguistics*, COLING-92, Nantes, vol. 3/4, pp. 1069-1073.

PHAN, H. K. et C. BOITET (1992) : « Multilinguization of an Editor for Structured Documents. Application to a Trilingual Dictionary », C. Boitet (dir), *Proceedings of the 14th International Conference on Computational Linguistics*, COLING-92, Nantes, vol. 3/4, pp. 966-971.

RICHARDSON, S. D. (1985) : *Enhanced Text Critiquing Using a Natural Language Parser : the CRITIQUE System*, RC 11332, IBM, Thomas J. Watson Research Center, Yorktown Heights.

RICHARDSON, S. D. et L. C. BRADEN-HARDER (1988): « The Experience of Developing a Large-scale Natural Language Text Processing System: CRITIQUE », *Proceedings of the 2nd Conference on Applied Natural Language Processing.*

RIMON, M., MCCORD, M. C., SCHWALL, U. et P. MARTÇNEZ (1991) : « Advances in Machine Translation Research in IBM », *Proceedings MTS-III (MT Summit)*, Boston, 1-4 July 1991, 11-18.

ROLLING, L. (1990) : « Trends of Multilingual Machine Translation in Europe », *Proceedings MMT'90*, Tokyo, 5-6 Nov. 1990, 2 p.

SADLER, V. (1989) : « Working with Analogical Semantics: Disambiguation Technics in DLT », T. Witkam (dir), *Distributed Language Translation (BSO/Research)*, Floris Publications, Dordrecht, Holland, 256 p.

SIGURDSON, J. et R. GREATEX (1987) : *MT of On-line Searches in Japanese Data Bases*, RPI, Lund Univ., 124 p.

SOMERS, H., TSUJII, J. et D. JONES (1990) : « Machine Translation Without a Source Text », H. Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, vol. 3, pp. 271-276.

TOMITA, M. (1984) : « Disambiguating Grammatically Ambiguous Sentences by Asking », *Proceedings of the 10th International Conference on Computational Linguistics*, COLING-84, Stanford, 2-6 juillet 1984, ACL, pp. 476-480.

TOMITA, M. (1985) : « Feasibility Study of Personal/Interactive Machine Translation System », S. Nirenburg (dir), *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-85, Hamilton, N.Y., 14-16 Aug. 1985, Colgate University, vol. 1/1, pp. 289-297.

TOMITA, M. (1986) : « Sentence Disambiguation by Asking », *Computers and Translation*, 1/1, pp. 39-51.

VASCONCELLOS, M. et M. LEÓN (1988) : « SPANAM and ENGSPAM : Machine Translation at the Pan American Health Organization », J. Slocum (dir), *Machine Translation Systems*, Cambridge University Press, pp. 187-236.

VAUQUOIS, B. (1988) : *BERNARD VAUQUOIS et la TAO, vingt-cinq ans de Traduction Automatique, ANALECTES. BERNARD VAUQUOIS and MT, twenty-five years of MT*, C. Boitet (dir), Ass. Champollion et GETA, Grenoble, 700 p.

VAUQUOIS, B. et C. BOITET (1985) : « Automated Translation at Grenoble University », *Computational Linguistics*, 11-1, pp. 28-36.

VAUQUOIS B. et S. CHAPPUY (1985) : « Static Grammars: A Formalism for the Description of Linguistic Models », S. Nirenburg (dir), *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-85, Hamilton, N.Y., 14-16 Aug. 1985, Colgate University, pp. 298-322.

VERONIS J. et N. IDE (1990) : « Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine-Readable Dictionaries », H. Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, pp. 389-394.

WEAVER, A. (1988) : « Two Aspects of Interactive Machine Translation », M. Vasconcellos (dir), *Technology as Translation Strategy*, State University of New York at Binghamton, Binghamton, pp. 116-123.

WEHRLI, E. (1990) : « STS : An Experimental Sentence Translation System », H. Karlgren (dir), *Proceedings of the 13th International Conference on Computational Linguistics*, COLING-90, Helsinki, vol. 1/3, pp. 76-78.

WEHRLI, E. (1991) : « Pour une approche interactive au problème de la traduction automatique », *L'environnement traductionnel. La station de travail du traducteur en l'an 2001*, Actes du Colloque de Mons (Belgique), actualité scientifique, Sillery, AUPELF-UREF, Presses de l'Université du Québec, pp. 59-68.

WEHRLI, E. (1992) : « The IPS System », C. Boitet (dir), *Proceedings of the 14th International Conference on Computational Linguistics*, COLING-92, Nantes, vol. 3/4, pp. 870-874.

WHITELOCK, P. J., WOOD, M. M., CHANDLER, B. J., HOLDEN, N. et H. J. HORSFALL (1986) : « Strategies for Interactive Machine Translation: the Experience and Implications of the UMIST Japanese Project », *Proceedings of the 11th International Conference on Computational Linguistics*, COLING-86, Bonn, 25-29 août 1986, IKS, pp. 25-29.

WILKS, Y. (1973) : « An Artificial Intelligence Approach to Machine Translation », Shank et Colby (dir), *Computer Models of Thought and Language*, Freeman et Co, pp. 114-151.

WOOD, M. M. (1989) : « Japanese for Speakers of English: The UMIST/Sheffield Machine Translation Project », J. Peckham (dir), *Recent Developments and Applications of Natural Language Processing*, Kogan Page ltd, London, pp. 56-64.

WOOD, M. M. G. et B. CHANDLER (1988) : « Machine Translation For Monolinguals », D. Várgha (dir), *Proceedings of the 12th International Conference on Computational Linguistics*, COLING-88, Budapest, 22-27 Aug. 1988, ACL, pp. 760-763.

ZAHARIN, Y. (1986) : *Strategies and Heuristics in the Analysis of a Natural Language in Machine Translation*, Ph.D. thesis, Universiti Sains Malaysia, Penang (research conducted in cooperation with GETA, Grenoble).

ZAJAC, R. (1988) : « Interactive Translation: A New Approach », D. Várgha (dir), *Proceedings of the 12th International Conference on Computational Linguistics*, COLING-88, Budapest, 22-27 Aug. 1988, ACL.

ZEMB, J.-M. (1982) : « Les occurrences phématiques, rhématiques et thématiques des archilexèmes *modaux*. La notion sémantico-logique de modalité », *Recherches Linguistiques VIII*, Université de Metz et Klincksieck, pp. 75-116.

9

Transfert de la langue parlée japonais-anglais dans le système de traduction automatique ASURA

Mutsuko TOMOKIYO

ATR Interpreting Telecommunications Research Laboratories, Hikaridai, Seika-cho, Kyoto, Japon

• *Abstract* •

In ATR's ASURA system of speech translation, the E-J transfer is performed with rewriting rules specified to English-Japanese transfer. Certain rewriting rules concern the IFT (illocutionary force), which is a kind of language-independent label. The IFT is located at the top level in the semantic representation of a sentence. Almost all rewriting rules are defined with special features `in` *and* `out`*.* `in` *represents a logico-semantic pattern which should match the incoming logico-semantic representation of the source text.* `out` *represents the target one to be output. When pattern matching succeeds, the logico-semantic features of the source language are rewritten into the logico-semantic features indicated in the* `out` *portion. The majority of the transfer rules handles the semantic, syntactic and pragmatic differences between the two languages, in the context of a whole sentence.*

Introduction

Notre projet a été lancé pour sept ans en 1986 par le gouvernement japonais pour étudier les questions fondamentales que pose la TA de la parole, en japonais-anglais (J-A) et anglais-japonais (A-J). Un nouveau projet de sept ans dans le même domaine a commencé en avril 1993.

Nous avons construit un prototype J-A et A-J d'interprétation simultanée de dialogues oraux sur le réseau téléphonique. L'objet du dialogue est d'obtenir des renseignements sur des conférences internationales. Les interlocuteurs sont supposés être un(e) demandeur(se) et un secrétaire de la conférence.

La TA du dialogue oral s'effectue en trois phases successives : reconnaissance de la parole, la TA proprement dite, et synthèse de la parole. La TA est effectuée en trois étapes : analyse, transfert et génération.

Dans cet article, nous nous limitons à l'étape du transfert.

Le système entier, appelé ASURA, utilise trois formalismes différents pour l'analyse, le transfert et la génération. L'analyse est effectuée par un analyseur à cartes (*chart parser*) qui interprète une grammaire de quasi-unification. Le transfert et la génération sont fondés sur deux « moteurs » différents, qui interprètent des règles de réécriture. Ces deux moteurs utilisent l'unification, ainsi que des primitives de reconnaissance des formes (*pattern-matching*) et de contrôle. La grammaire linguistique de notre travail est une grammaire HPSG qui fournit des représentations structurales adéquates. À toutes les étapes, la représentation de l'énoncé contient une trichotomie de syntaxe, sémantique et pragmatique, encodée dans une seule structure de traits.

Le propos de cet article est de présenter les règles de réécriture et de décrire une affectation d'un « trait » particulier contenu dans les f-structures, appelé le *type de force illocutoire*, ou IFT. Nous exposerons, dans un premier temps, la démarche de notre système de réécriture. Nous décrirons ensuite l'affectation de l'IFT et sa vérification linguistique. Finalement, nous présenterons un résultat de transfert et génération.

D'une part, la plupart des règles de réécriture utilisent les traits spéciaux `in` et `out`. `in` représente la structure de traits logico-sémantiques de la langue source obtenue en fin d'analyse. `out` représente la structure de traits logico-sémantiques de la langue cible. Dans la règle de réécriture, `out` porte forcément une étiquette, appelée *illocutionary force type* (IFT), qui est indépendante de la langue source et cible. L'IFT est déterminé en fonction des propriétés grammaticales, de la modalité, de la classe des verbes principaux de l'énoncé, etc. D'autre part, le transfert de traits pragmatiques change certains paramètres communicatifs tels que le niveau de langue qui est radicalement différent entre le japonais et l'anglais.

Écrire des règles de réécriture, c'est formaliser les différences de représentation structurale entre deux langues. Essentiellement, le travail du transfert consiste à transformer la représentation structurale de la langue source en celle de la langue cible.

La différence entre les deux langues parlées sera présentée du point de vue linguistique, et des phénomènes variés seront illustrés en prenant des exemples réels de la langue source et de la langue cible.

La qualité du transfert fondé sur l'IFT dépend impérativement de l'exactitude de l'IFT choisi dans le résultat d'analyse.

L'avantage de cette méthode est qu'il est possible de neutraliser les différences syntaxiques entre deux langues au niveau du transfert, et qu'il est possible de s'occuper de l'autre langue systématiquement et indépendamment.

Le problème est qu'il est difficile de déterminer la valeur de l'IFT à partir du résultat d'analyse. Par exemple, la négation grammaticale n'est pas nécessairement une

expression négative du point de vue de l'IFT, et l'interrogatif grammatical n'exprime pas forcément une question adressée à l'interlocuteur.

Pour affecter une seule valeur à l'IFT, les détails pertinents de l'énoncé doivent être considérés. Nous discuterons le problème de l'affectation de l'IFT, en mettant en contraste les langues anglaise et japonaise.

Au cours de ce travail, nous avons pu constater que, dans la plupart des cas, les problèmes et les solutions en TA se transposeraient aisément au japonais-français.

Différences entre l'anglais et le japonais

Le travail du transfert est de prendre en compte les différences linguistiques entre les deux langues. Essayons de définir les différences syntaxiques et sémantiques du point de vue de la TA. Les différences syntaxiques sont les suivantes :

(1) Le mot-clé de la structure de phrase se situe à gauche dans une phrase anglaise, mais à droite en japonais
Par exemple : ***books*** *on linguistics*
言語学 の 本

(2) Il n'y a pas de *WH-movement* dans les phrases inrerrogatives japonaises, ni de déplacement du sujet.
Par exemple : ***What*** *is the topic of the conference?*
会議の テーマは 何ですか

(3) L'ordre des éléments de phrase est fixé en anglais, contrairement à celui du japonais qui est relativement libre.
Par exemple : *There is a book on the desk.*
机の上に 本が あります / 本が 机の上に あります

(4) Il n'y a pas d'explétif en japonais ; il est obligatoire en anglais.
Par exemple : *There is a book on the desk.*
机の上に 本が あります。

(5) Il peut exister plusieurs sujets dans une phrase japonaise, alors que c'est impossible en anglais.
Par exemple : 文明国が 男性が 平均寿命が 短い (kuno 1970)
civilized countries, male, the average lifespan is short. (It is civilized countries that man, their lifespan is short in.)

(6) Le pronom ne peut être modifié en anglais, mais bien en japonais.
Par exemple : * *last week's he*
先週の 彼

Nous traitons la différence concernant l'ordre des éléments de phrase dans une

phrase (1,3) lors de l'étape de génération. Mais des phénomènes linguistiques tels que les sujets multiples ne sont pas traités. Le reste est traité à l'étape de transfert.

Les différences sémantiques entre les deux langues sont les suivantes :

(1) Le japonais met l'accent sur la structure *thème-rhème*. L'anglais le met sur la structure *sujet-prédicat.*
Par exemple : *Where is the conference venue?*
会議場 は どこですか

(2) L'*honorifique* japonais est représenté *grosso-modo* par les lexies de l'*honorifique.* L'interlocuteur montre sa révérence envers une personne de plus haute classe sociale, en utilisant les lexies de respect ou d'humilité. L'anglais est stratégiquement varié, quoique plus systématique dans l'usage de l'*honorifique.*
Par exemple : *Will your wife attend the conference?*
奥様 は 会議に ご 出席 になり ますか

(3) L'ellipse du sujet se produit fréquemment dans l'énoncé japonais, contrairement à l'anglais où il se produit rarement.
Par exemple : *Will you attend the conference?*
会議に ご出席になりますか

(4) Le système du déictique japonais est plus riche que celui de l'anglais, parce qu'il y a une productivité par concaténation de verbes.
Par exemple :買って あげます
I'll buy it for you.
Il n'y a pas de sujet, ni d'objet dans la phrase japonaise. Cependant, la relation entre les interlocuteurs peut être interprétée comme celle de mère à enfant à cause du système du déictique.

(5) La notion de singulier et de pluriel grammatical existe implicitement en japonais, alors qu'elle existe explicitement en anglais.
Par exemple : *Close your **eyes**.*
目 を つむってごらん
En japonais, le concept des yeux est pluriel. Il faut indiquer œil gauche, œil droit ou un seul œil, si nous voulons faire fermer un seul œil à quelqu'un.

(6) La postposition japonaise sert à exprimer une relation grammaticale dans la phrase et en même temps une sensibilité inter-personnelle, alors que la postposition anglaise est une sorte d'ellipse du complément. La postposition japonaise a un rôle différent de la postposition anglaise : par exemple, dans (a), la postposition *on* peut être remplacée par une expression verbale, contrairement à la postposition のですが dans (b) qui représente une sorte de modalité d'énoncé et ne peut donc être remplacée par rien.
(a) *They travelled **on**.* → *They **continued** their journey.*
(b) 会議に 参加したい のですが (*I would like to attend the conférence.*)

(7) L'énoncé japonais consiste en des groupes adverbiaux dont la relation

grammaticale n'est pas nécessairement claire, contrairement à l'énoncé anglais qui consiste en des propositions coordonnées et subordonnées.
Par exemple : *I live in Tokyo and will go to Osaka two days before the conference for some business. I want to head to the conference after the business, but as the schedule is pushing, I want to spend my free time in the conference for my business. I'm now adjusting my schedule, though.*
実は、私 、東京におりますもんですから、 会議の二日前に大阪の方に伺ってですね、そちらでちょっと仕事片付けた後、会場の方に伺わせていただいて、会議に参加しようと思っているんですけども、 なにぶんちょっと仕事の方が溜まっておりますもんですから、 会議の期間中もですね、できましたら 暇がありましたら 自分の仕事に充てたいと思っておりまして、 今そういった日程の方ですね、調整してるところなんですけども。
L'énoncé japonais a dix propositions pour une phrase, contrairement à l'énoncé anglais constitué de trois phrases séparées.

(8) La façon japonaise de parler montre clairement à qui appartiennent initialement les renseignements en question.
Exemple : A- 東京で 学会が あるよ (*It seems that a conference will be held in Tokyo.*)
B-その 学会は 何というの (*What is the name of **the** conference?*)
A-この学会の名前は プラグマティックスというんだ(***This** is called the Pragmatics conference.*)

Le thème de la conférence était introduit par A dans la conversation, qui met le déterminant この (*this*) avant *conférence*. Le déterminant employé par B est その (*that*).
Les Japonais ont tendance à se servir d'expressions atténuées.
Exemple : ***I'd like to** attend the conference.*
会議に 申し込みたい のですが

Parmi les différences sémantiques, l'ellipse du sujet est résolue lors de l'étape de génération. Le reste est traité à l'étape de transfert.

Système de réécriture

Nous donnons une explication simple concernant le système de réécriture et les règles de réécriture.

Démarche du système de réécriture

Le système de réécriture (RWS) est un système dans lequel nous réécrivons les f-structures avec des règles définies par une formule unique. Les règles sont constituées d'un index, d'une définition et d'un corps de règle.

```
(rew :defrwschema2 def528 V EX                                    index
" on < OBJE RELN> be-vi-5 in :phase :E-J :type :default

in = [[RELN unknown-IFT]                                      definition
      [AGEN ?agen]
      [RECP ?recp]                                               body
      [OBJE [[RELN be-vi-5]
             [OBJE ?obje]
             [IDEN ?iden]
              ?rest]]]

SET PARAMETER :IFT :INFORM

out = [[RELN inform]
       [AGEN ?agen]
       [RECP ?recp]
       [OBJE [[RELN desu-aux-1]
              [OBJE ?obje]
              [IDEN ?iden]
               ?rest]]]                                              (RW.1)
```

Une règle définit :

- le chemin des traits : `OBJE RELN`
- le nom des traits : `EXPECT-VT-1`
- une condition d'application pour la règle : `in :PHAS :E-J :TYPE :default`.

Le corps de règle est constitué de `in` et `out`. `in` représente la structure de traits logico-sémantiques de la langue source obtenue en fin d'analyse. `out` représente la structure de traits logico-sémantiques de la langue cible. Dans la règle de réécriture, `out` porte forcément une étiquette, appelée IFT, qui est indépendante des langues source et cible. Les f-structures obtenues en fin d'analyse sont comparées aux parties `in` des règles actives, ces parties étant interprétées comme des schémas de f-structures. Quand le *pattern-matching* réussit, les f-structures concernées sont remplacées par les f-structures de `out`. Les f-structures de `out` sont celles de la langue cible qui correspondent à la langue source.

Règles de réécriture

Il y a trois différents genres de règles : les règles fonctionnelles, les règles grammaticales et les règles lexicales. Une règle fonctionnelle :

- prépare un environnement dans lequel la règle d'IFT sera acceptée ;
- établit une situation conversationnelle ;
- indique l'ordre et la façon dont les règles sont appliquées.

Une règle de grammaire s'occupe de réécrire le temps, l'aspect, la négation, le passif, les substantifs verbaux, etc.

Une règle lexicale réécrit les lexies anglaises en japonais.

La valeur de RELN, UNKNOWN-IFT dans RW.1 est une étiquette d'IFT fonction de SET PARAMETER : IFT. AGEN et RECP sont remplacés respectivement par SPEAKER et HEARER parce que, dans notre domaine de travail, deux interlocuteurs se parlent au téléphone.

La représentation sémantique produite par l'analyse est constituée d'une partie sémantique et d'une partie pragmatique. Les règles concernant la pragmatique sont codées de la même façon que celle concernant la sémantique.

La f-structure sémantique contient le nom des relations (RELN), le temps, l'aspect, les substantifs verbaux. La f-structure pragmatique contient la topicalisation, la présupposition, la politesse, etc.

Réduction de la *force illocutoire* au trait IFT

La force illocutoire que nous considérons est la force de performance potentielle des énoncés à travers l'activité de la langue. Elle provoque des réactions de l'auditeur dans le dialogue. En conséquence, elle est classifiée en groupes supposant des réactions dialogiques différentes de l'auditeur. Chaque groupe porte un nom, appelé un IFT. La détermination finale de l'IFT est faite en fonction de propriétés grammaticales, de la modalité, de la classe des principaux verbes de l'énoncé, etc.[1]

L'IFT prend la forme d'une étiquette par phrase et sert à déterminer le niveau structurel de la phrase pour la phase de génération.

Force illocutoire

Nous supposons que la phrase peut être analysée en une proposition contenant un verbe de performance et des propositions propositionnelles (*propositionnelles*). Une propositionnelle est supposée enchâssée dans la proposition performative, bien qu'elle soit en général supprimée dans la conversation réelle.

La *force illocutoire* d'un énoncé dépend du verbe de performance et de la propositionnelle.

1. Le module d'inférence n'est pas encore implanté pour le discours dans notre système.

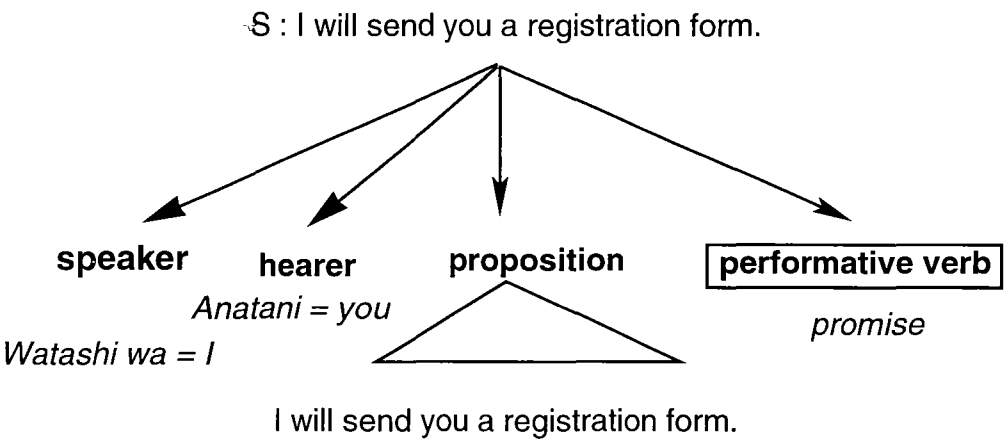


FIGURE 2 : Watashiwa anatani tourokuyousiwo okurimasu.

Nous appelons verbe de performance des verbes tels que *promise* (voir fig.2) et propositionnelle des propositions telle que *I will send you...* (toujours fig.2). D'une part, l'IFT est affecté en considérant de façon synthétique le verbe performatif, la propositionnelle et certaines propriétés grammaticales. Observer la grammaire de la phrase sert à lever l'ambiguïté parmi les IFT possibles. D'autre part, l'IFT n'est valable que si une condition, appelée condition préparatoire (Searle et Bierwisch 1980), est vérifiée. La condition préparatoire limite l'environnement de la conversation : dans notre domaine de travail, un dialogue téléphonique, les interlocuteurs sont supposés être un(e) demandeur(se) et un(e) secrétaire de conférence.

Propriétés grammaticales

Pour affecter un IFT à une phrase, la phrase est d'abord identifiée comme étant énonciative, interrogative, impérative, causative ou bien passive. Deuxièmement, elle est identifiée comme étant affirmative ou négative. Troisièmement, elle est identifiée comme étant indicative, subjonctive ou bien conditionnelle. Finalement, nous examinons le temps, l'aspect et le sujet de la phrase.

Proposition propositionnelle

La propositionnelle définie ici est le contenu de ce que le locuteur veut dire à l'auditeur à travers la conversation. Elle est classifiée en groupes, en considérant le sens essentiel du verbe de la propositionnelle (voir Leech et Svartvik 1975 ; Tomokiyo et Morimoto 1992).

> *statement, command, obligation, agreement-disagreement, fact-hypothesis, tentative-open condition, possibility-impossibility, ability-inability, certainty-uncertainty, probability-improbability, belief-assumption, liking-disliking, appearance, likeness, hearsay, volition, reason, understatement*

Proposition performative et ses verbes

La proposition performative définie ici est la phrase du verbe[2] de performance dans laquelle le sujet est le locuteur, 1re personne du singulier = *I*, et dans laquelle le complément d'objet indirect est l'auditeur, 2e personne du singulier = *you*, comme nous l'avons vu dans la figure 2. La phrase performative est l'hyper-nœud au niveau structurel de la phrase. En même temps, elle est, au niveau grammatical, la proposition principale suivie par la propositionnelle comme complément d'objet direct.

La plupart des verbes performatifs sont supprimés dans une conversation réelle. Dans ce cas-là, ils doivent être rétablis pour l'affectation de l'IFT. Par contraste, peu de verbes et auxiliaires tels que *would like to, promise, advise,* etc. apparaissent dans les énoncés (voir Yasui *et al.*). Le calcul de notre IFT dépend donc largement de notre capacité à établir les verbes performatifs absents.

La phrase de la figure 2 est une phrase énonciative, affirmative et indicative conformément à notre classification. Le temps et l'aspect de la propositionnelle sont respectivement *future et unreal.* Le verbe de la propositionnelle est jugé *fact.* Ainsi, le verbe performatif rétabli est *promise*[3]. La structure interne de la phrase est donc : *I promise you that I will send you a registration form.* De cette façon, *promise* est affecté comme IFT de la phrase. Voyons un autre exemple :

a. *Please tell me about the topic of the conference.*
b. *I would like to apply for the conference.*
c. *May I have your name and your address?*

compendium grammatical	phrase a.	phrase b.	phrase c.
énonciative		*yes*	
interrogative			*yes*
impérative	*yes*		
causative			
passive			
indicative	*yes*	*yes*	*yes*
subjonctive			
conditionnelle			

2. Les verbes de performance contiennent des auxiliaires.
3. Pour des exemples de classification des verbes de performance, voir R. Lakoff, McCawley, Ross, Searle. *order, command, request, ask, beg, demand, invite, prohibit, advise, recommand, suggest, councel, thank you, permit, allow, declare, claim, insist, name, offer, would like to, may, etc.*

affirmative	*yes*	*yes*	*yes*
négative			
temps	*present*	*present*	*present*
aspect	*unreal*	*unreal*	*unreal*
verbe de performance		*would like to*	*may*
propositionnelle	*statement*	*fact*	*fact*
adverbe	*politeness*		
sujet	*2nd*	*1st*	*1st*
wh-éléments			
forme contractée			
symbole			?

(list 1)

(list 1) donne une comparaison des propriétés grammaticales pour les trois phrases : quel est le verbe de performance, quel genre de verbe se trouve dans la propositionnelle, et quel sujet est contenu dans la phrase ? Après avoir fait des pointages grammaticaux, *request* est choisi comme IFT de ces trois phrases différentes.

Au total, nous avons déterminé 12 traits d'IFT. L'IFT est fixé par des conditions propres :

request[4] : demande du locuteur à l'auditeur [[CAT[5] s][SMOD interrogative][TEMPS present][PV[6] may]]
questionif : questions auxquelles une des deux réponses, *oui* ou *non*, est possible[[CAT s][SMOD indicative, interrogative][SYMBOL ?]]
questionref : questions auxquelles n'importe quelle réponse peut être donnée à moins que le locuteur ne donne des renseignements par des pronoms interrogatifs [[CAT s][SMOD indicative][WH (:set why what where how who)][SYMBOL ?]]
tag-question : question attachée en fin de phrase pour confirmer la vérité de l'énoncé [[CONTRACTED yes][SYMBOL ?]]
echo-question : question par laquelle l'auditeur demande au locuteur de répéter une

4. Nous montrerons un des cas différents pour un IFT.
5. CAT signifie catégorie.
6. PV signifie verbe de performance.

information, en général parce qu'il l'a mal ou pas entendue[7]. [[CAT NOT: s] [SYMBOL?]]
responseif : réponse à une questionif [[CAT adv]]
responseref : réponse à une question par un pronom interrogatif [[CAT NOT: adv]]
emotional-response : pour montrer son intérêt, sa surprise, son plaisir, etc.[8] [[CAT (set: adv inter idiom)][SYMBOL ?]]
promise : le locuteur promet à l'auditeur de faire quelque chose [[CAT s][TEMPS (set:present future)][SMOD NOT:subjonctive] [SUBJ 1st]]
suggest : le locuteur suggère à l'auditeur de faire quelque chose[9]. [[CAT s][PROP[10] obligation]] or [[CAT s][VFORM inf][SUBJ 2st]]
inform : énoncé où le locuteur fournit à l'auditeur des informations [[CAT s][SMOD NOT:imperative]]
invitation : le locuteur propose à l'auditeur de faire quelque chose ensemble [[CAT s] [SMOD imperative][PV let's][TEMPS present]] or [[CAT s][SMOD interrogative][PROP invitation][TEMPS present] [SUBJ 1st]]
phatic : salutations [[CAT interj]]
expressive : le locuteur manifeste son sentiment [[CAT idiom][PV expressives]]
encouragement : Le locuteur montre qu'il continue d'écouter. [[CAT adv]]

Vérifications linguistiques

Nous donnons ici des explications linguistiques au sujet de la détermination de l'IFT.

IFT et personne grammaticale

L'IFT dépend de la personne grammaticale. Ce qui suit indique que le sujet de la propositionnelle influe sur l'affectation de l'IFT.

(a) *I would like to apply for the conference. (request)*
(b) *You (He, She, They) would like to apply for the conference.*

En (a), le locuteur voudrait s'inscrire à la conférence, contrairement à (b) où il parle pour quelqu'un d'autre. En japonais, (a) est représenté par une expression atténuée de demande, たいのですが (*tainodesuga*) alors que (b) est représenté par une expression de présomption, がっている (*gatteiru*). C'est pour cela que nous examinons le sujet de la propositionnelle lors de l'affectation de l'IFT à une phrase.

À propos de *would like to*, deux règles différentes sont appliquées. Les f-structures obtenues par l'analyse sont d'abord transférées en *1st-person-wish* en regardant le sujet de la phrase, puis *1st-person-wish* est réécrit en *request* dans l'IFT.

---

7. Par exemple : *I didn't enjoy that meal. – Did you say you didn't enjoy it?*
8. Par exemple : *I heard Paula's getting married. Really?*
9. Par exemple : *They suggested Smith shoud be dropped from the team. They suggested that Smith be dropped from the team.*
10. PROP signifie le contenu de la proposition.

règle : *would like to*

```
in= [[RELN UNKNOWN-IFT]
     [AGEN ?AGEN]
     [RECP ?RECP]
     [OBJE [[RELN WOULD_LIKE_TO-1]
            [OBJE ?OBJE]
            [EXPR [[RELN I-PRON-1]]]
                                ?rest]]]
out= [[RELN UNKNOWN-IFT]
      [AGEN ?AGEN]
      [RECP ?RECP]
      [OBJE [[RELN 1ST_PERSON_WISH]
             [OBJE ?OBJE]
             ?rest]]]      (RW.3)
```

IFT et *polarité*

L'IFT dépend de la *polarité* du prédicat de la phrase. Par exemple, deux IFT différents sont affectés aux deux phrases suivantes, bien qu'elles utilisent toutes deux *would like to.*

(a) *I would like to apply for the conference. (request)*
(b) *I wouldn't like to apply for the conference. (inform)*

En (a), le locuteur voudrait provoquer une réaction de quelqu'un dans son intérêt alors qu'il dit ce qu'il pense en (a). En japonais, (a) est représenté par une expression atténuée de demande, par contraste avec (b), représenté par une expression neutre, たくありません (*takuarimasen*). C'est pourquoi nous faisons référence à la *polarité* d'une phrase pour déterminer l'IFT.

IFT et *pragmatique*

Des phénomènes linguistiques tels que la *présupposition*, les expressions logiquement substituées ou le *sous-entendu* interviennent aussi (Sgall, Hajičová et Panevova 1986).

Par exemple : *He will also come. (Someone will come and he will also come.)* → *Karemokuru.*

En japonais, la plupart des présuppositions sont représentées par des particules spéciales. La présupposition doit donc être explicitement spécifiée. Elle est mise en œuvre dans la *pragmatique* au cas où un mot-clé qui dévoile la présupposition existe dans la langue source. La réécriture de la pragmatique est effectuée de la même façon que celle de la sémantique.

Les sous-entendus dépendant du contexte ou de conventions sociales ne sont pas encore traités dans notre système.

Par exemple : *It's very cold here.* → *Will you close the door, please?*

IFT et spécificité du prédicat

L'IFT dépend de ce qu'une phrase est marquée ou non spécialement par l'interrogative ou la comparative. L'interrogative est, en général, une phrase non marquée sous forme affirmative. Ce qui suit indique que la forme négative-interrogative influe sur l'affectation du l'IFT.

(a) *Why don't you send me a registration form? (request ou complaint)*
(b) *Why do you send me a registration form? (question)*

En (a), le locuteur se plaint de ce que l'auditeur ne lui a pas envoyé un formulaire. En japonais, on utilise une expression de faveur négative,てくれない (*tekurenai*). En (b), le locuteur demande pourquoi l'auditeur lui a envoyé un formulaire. En japonais, on utilise une interrogative neutre. L'IFT doit donc être différent pour l'une et pour l'autre. À propos de la comparative, le prédicat adjectival de la phrase influe sur l'affectation de l'IFT.

(a) *Paul is taller than Michel.*
(b) *Paul is shorter than Michel.*

(a) est interprété comme *Paul est plus grand que Michel,* mais nous ne savons pas pourtant s'ils sont tous grands car, dans le couple (*tall/short*), *tall* est neutre et *short* marqué. (b) signifie en plus qu'ils sont beaucoup plus petits que les autres. C'est une conséquence du fait que les phrases portent la polarité de l'adjectif. Les sous-entendus de ce genre ne sont pas traités dans notre système.

IFT et *prosodie*

L'IFT dépend aussi largement de la prosodie des énoncés. Cependant, pour l'instant nous ne la traitons pas dans notre système (Leech et Svartvik 1975 ; Halliday 1989).

Conclusion

Nous avons montré dans cet article de quelle façon nous déterminons les valeurs d'IFT et comment cela nous permet de transférer la modalité en même temps que la sémantique de l'anglais au japonais. Le transfert reposant sur l'IFT s'avère une méthode souple et puissante qui permet de neutraliser les différences syntaxiques radicales qui existent entre deux langues comme le japonais et l'anglais.

Le traitement du discours linguistique et d'énoncés complexes s'appuyant sur l'IFT est une perspective prometteuse. Pour le moment, nous ne traitons que des énoncés simples qui contiennent deux ou trois propositions au maximum. L'IFT ne prend la forme que d'une étiquette par phrase. Mais, d'une part, la conversation réelle est divisée en unités d'énoncés plus longues, et d'autre part, il reste à résoudre l'ambiguïté d'IFT, qui est impliquée dans les résultats du transfert : par exemple, entre *response* et *encouragement.* Cette ambiguïté doit être résolue de façon à traiter le discours du texte. Si le système de transfert nous permettait de noter l'histoire de l'IFT, c'est-à-dire l'IFT de la phrase précédente, un calcul utilisant cette information pourrait lever certaines de ces ambiguïtés.

Résultat du transfert et de la génération

Nous montrons la succession du transfert et de la génération dans une phrase exemple. Supposons qu'une règle d'IFT est appliquée à la phrase suivante :

I will send you a registration form. (promise)
登録用紙を お送りします*(Tourokuyousiwo Ookurishi**masu**)*[11]

Le résultat de transfert est :

```
[[SEM [[RELN PROMISE]
       [AGEN !X1[[LABEL *SPEAKER*]]]
       [RECP !X2[[LABEL *HEARER*]]]
       [OBJE  [[RELN 送る-V-1]
               [TENSE FUTURE]
               [[AGEN !X1]
                [INDEX [[PERSON 1st]]
               [RECP !X2]
               [OBJE [[PARM !X3[]]
                      [RESTR [[RELN 登録用紙 -1]
                              [ENTITY !X3]]]]]
               [SEM-ASPE UNREAL]]]]]
[PRAG [[RESTR [[IN [[FIRST [[RELN POLITE]]]
                   [REST [[FIRST [[RELN CONDESCEND]]]
                          [REST [[FIRST [[RELN POLITE]]]
                                 [REST [[FIRST [[RELN POLITE]]]
                                        [REST !X5[]]]]]]]]]]
              [OUT !X5]]]
        [BENEFIT HEARER-SIDE]
        [HEARER !X2]
        [INTENTION OFFER]
        [PRSP-TERMS [[PRSP-MOD NULL]]]
        [SPEAKER !X1]
        [TOPIC [[TOPIC-MOD NULL]]]]]]
```

La règle de génération appliquée à la f-structure obtenue en fin de transfert, est la suivante :

```
(g_pd
 : name S2PROMISE-S1_FP_SIGN
 : internal _structure (S2  S1  FP  SIGN)
 : annotation
 ((S2   [[syn [[cat s2]]]
         [sem [[reln promise]
               [agen [[label *speaker*]]]
               [recp [[label *hearer*]]]
               [obje ?proposition]]]])
 (S1    [[syn [[cat s1]]]
```

11. (b) *I will tell you about a registration form. (inform) (Tourokuyousnituite ohanashishi**masyou**)*

```
        [sem ?proposition]])
 (FP    [[syn [[cat fp]]]
        [sem ます]])
 (SIGN [[syn [[cat sign][lex "."]]]])))
```

La structure arborescente de sortie de la génération est alors :

```
S2 [#762 EXPANDED (S2PROMISE-S1_FP_SIGN)]
¦--S1  [#763  EXPANDED  (S1-S  S-NP_VP  VP-NP_VP1  VP1-NP_V
V-OKURU_ASPL)]
¦   ¦--S [#764 EXPANDED NIL]
¦      ¦--NP [#765 EXPANDED (NP-N_P2 N-SPEAKER)]
¦      ¦  ¦--N [#766 LEXIFIED "*speaker*" NIL]
¦      ¦  ¦--P2 [#767 LEXIFIED "wa" (WA)]
¦      ¦--VP [#768 EXPANDED NIL]
¦         ¦--NP [#769 EXPANDED (NP-N_P4 N-HEARER)]
¦         ¦  ¦--N [#770 LEXIFIED "*hearer*" NIL]
¦         ¦  ¦--P4 [#771 LEXIFIED "ni" (NI)]
¦         ¦--VP1 [#772 EXPANDED NIL]
¦            ¦--NP [#773 EXPANDED (NP-N_P3 N-TOUROKUYOUSHI)]
¦            ¦  ¦--N  [#774 LEXIFIED "tourokuyoushi" NIL]
¦            ¦  ¦--P3 [#775 LEXIFIED "wo" (WO)]
¦            ¦--V [#776 LEXIFIED "okuri" NIL]
¦--FP1 [#777 LEXIFIED "masu" (FP1_MASU)]
¦--SIGN [#778 LEXIFIED "." NIL]
```

Références

HALLIDAY, M. A. K. (1989) : *Spoken and Written Language*, Oxford University Press.

HASEGAWA, Toshiro (1990) : *Feature Structure Rewriting Manual*, TR-I-0187 ATR.

HASEGAWA, Toshiro (1990) : *Details of the Transfer Process in the ASURA Translation System*, TR-I-0188 ATR.

LARREYA, Paul (1979) : *Énoncés performatifs présupposition*, Nathan Université information formation.

LEECH, Geoffrey et Jan SVARTVIK (1975) : *A Communicative Grammar of English*, Londres, Longman.

SEARLE, J. R. et M. BIERWISCH (1980) : *Speech Act Theory and Pragmatics*, Dordrecht.

SGALL, Petr, HAJIČOVÁ, Eva et Jarmila PANEVOVA (1986) : *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*, D. Reidel Publishing Company.

TOMOKIYO, Mutsuko et Noriyoshi URATANI (1993) : *Japanese Generation within ASURA Framework*, TR-I-0349 ATR.

TOMOKIYO, Mutsuko et Tsuyoshi MORIMOTO (1992) : *Communicative Functions of Spoken Japanese and its Meaning Interpretation on MT System,* TR-I-0260 ATR.

YASUI, Minoru *et al.* (1983) : *Imiron*, Taisyuukansyoten 1983 Taiisyukansuten.

10

LEAF ou comment garder l'origine de l'ambiguïté

Mathieu LAFOURCADE

GETA, IMAG, Université Joseph-Fourier et CNRS, Grenoble, France

• *Abstract* •

This article presents a new abstract model, LEAF, which proposes a generical solution to ambiguity management in Natural Language Processing. This model uses a temporal (multilevel and weighted) lattice as factorizing structure. Linguistic decorations are worn by the lattice nodes. The lattice geometry and decoration are manipulated by linguistic tools. An architecture based on this model is proposed for written text interpretation. Nodes of the lowest level hold characters (potentially ambiguous). The morphological levels are subject to morphological processing. The third level is composed of derivation trees, decorated abstract trees and psi-terms (typed feature structures). It is possible to "unfold" an overlapped level to create a new level.

Introduction

Les différentes applications de traitement automatique de la langue naturelle (TALN) s'appuient généralement sur des formalismes linguistiques conçus pour trouver des solutions à quelques difficultés spécifiques de ce domaine ou pour fournir une approche plus confortable aux linguistes. Cependant, aucun formalisme n'arrive à gérer à lui seul l'ensemble des difficultés et ce qui est gagné sur un front est souvent perdu sur un autre. D'autre part, le problème de l'ambiguïté reste la pierre d'achoppement de la traduction automatique (et du TALN) mais peu a été fait pour sa représentation et moins encore pour la conservation de son origine. C'est à partir de ces deux constats que l'on souhaite proposer une méthodologie de programmation ambiguë permettant au linguiste de concevoir des applications linguistiques (à partir de langages spécialisés pour la programmation linguistique) soit en gérant les ambiguïtés de ma-

nière explicite, soit comme s'il n'y avait (à chaque instant) qu'une solution. Un tel style de programmation « non-déterministe » est fortement inspiré par des mécanismes universels comme ceux mis en place dans Prolog où la situation, déterministe en apparence, laisse place à une gestion automatisée du non-déterminisme. Un tel mécanisme se fonderait sur une structure factorisante et des opérateurs associés offrant la possibilité d'exprimer des solutions génériques au traitement de l'ambiguïté.

Le modèle LEAF (*Lattice, Engine And Features*) est la combinaison d'un treillis (temporel pondéré et multistrate), de systèmes de décorations linguistiques et d'outils linguistiques (applications de LSPL). Les décorations, calculées et manipulées pour les outils, sont portées par les nœuds du treillis. Ce treillis est par ailleurs modifié dans sa géométrie par ces mêmes outils. La multiplicité des strates permet la conservation des résultats d'analyses précédents. Cette approche non destructive, à la base de la possibilité de « mémorisation » des sources d'ambiguïté permet également d'envisager d'autres types de traitement. Entre autres, on pensera aux pondérations qui permettent une évaluation paresseuse ne calculant que les nœuds et les chemins les plus prometteurs. La visualisation pour le linguiste d'une structure de données aussi complexe peut prendre plusieurs formes et faciliter grandement la mise au point d'applications linguistiques. Ce modèle peut fournir une approche générique pour le TALN.

Après une présentation du modèle LEAF, cet article décrit les aspects d'une architecture spécifique pour l'interprétation des textes écrits. Cependant, certains choix sont clairement faits de manière à préserver une certaine généricité permettant éventuellement à d'autres types de traitements d'utiliser l'architecture proposée. Cette architecture est illustrée strate par strate. Les nœuds de la première strate contiennent les caractères et sont déjà potentiellement ambigus si l'entrée est le résultat d'un outil de reconnaissance optique de caractères ou de reconnaissance de l'écriture. Les strates morphologiques sont constituées d'une strate de catégories contenant des mots et d'une strate d'analyse morphologique contenant des informations beaucoup plus complètes. La strate linguistique est composée d'arbres concrets plongeant directement dans le treillis ainsi que d'arbres abstraits décorés contenus dans les nœuds. Des psi-termes (ou structures de traits typés) viennent décorer les nœuds des arbres abstraits. Il est possible à des fins de visualisation de « déplier » un niveau imbriqué, comme les arbres abstraits ou encore de représenter des nœuds de nature différente (instance de règles de programmes en LSPL) directement dans la strate linguistique, ou d'en faire une nouvelle strate.

Le modèle LEAF

LEAF est un modèle d'architecture logicielle destinée à apporter des réponses aux questions suivantes :

- Comment garder une trace de l'origine de l'ambiguïté ? On veut disposer d'un système simple permettant au linguiste informaticien de représenter et de gérer les ambiguïtés soit explicitement, soit implicitement.
- Quelle structure de données utiliser ? On souhaite disposer d'une structure générique indépendante d'un formalisme linguistique et pouvant être utilisée dans plusieurs domaines de TALN comme la reconnaissance de la parole, la reconnaissance de l'écriture et l'interprétation des textes écrits.

● Comment garantir la rapidité des traitements linguistiques ? Les domaines comme l'interprétation de textes écrits (ou la reconnaissance de l'écriture et la reconnaissance de la parole), intégrés dans le processus de communication homme-machine ne souffrent pas de délais trop importants. Cette rapidité est fonction de la richesse de représentation de la structure.

● Comment simplifier le travail du linguiste informaticien ? La visualisation des résultats d'analyse et la nature de l'interaction entre le linguiste et ces structures aura une influence sur la mise au point des applications linguistiques. L'accent est de plus en plus souvent mis sur l'aspect graphique dans la visualisation de structures de données très complexes (c'est, par exemple, le cas dans CYC, voir Lenat *et al.* 1990). Dans LEAF, la structure de treillis est facilement visualisable et il semble intéressant de pouvoir exhiber « à la demande » des nœuds les informations jugées pertinentes par l'utilisateur.

● Comment mettre en œuvre des techniques de génie linguiciel ? Un environnement de programmation linguistique serait dynamique, c'est-à-dire que la modification des attributs d'un objet, même au cours d'une analyse, a un effet immédiat sur l'état du système (pas de recompilation). La conception et la mise au point de programmes en serait grandement facilitées.

D'autre part, la conception d'un tel système est rendue plus modulaire. Il est plus aisé de faire communiquer plusieurs outils autour d'une seule structure de données partagée (technique du « tableau noir ») que de définir des protocoles *ad hoc* entre chaque catégorie d'outil (morphologique, syntaxique, etc.).

Présentation générale

Le modèle LEAF met en jeu trois types d'objets : un treillis, des décorations, des moteurs. Le modèle décrit les composantes générales de ces trois objets, ainsi que leur interrelation.

Ce modèle va servir de base à la définition de diverses architectures. Les descriptions détaillées des décorations et des moteurs restent à la charge des architectures implémentant ce modèle en vue d'applications particulières.

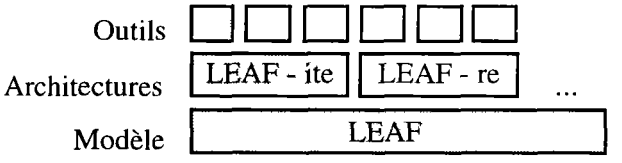


FIGURE 1 : Plusieurs architectures peuvent être définies à partir du modèle LEAF. Elles spécifient la composition des strates ainsi que les outils mis en jeu. Ces architectures peuvent être dédiées à des problèmes différents (interprétation de textes écrits, reconnaissance de l'écriture,...). Les outils définis se basent sur une architecture particulière.

Treillis

Les arcs liant les nœuds sont explicites et interdisent la création de cycles. Il existe

toujours un unique premier nœud, ⊢, et un unique dernier nœud, ⊣. Les nœuds portent des décorations et les arcs portent une valuation.

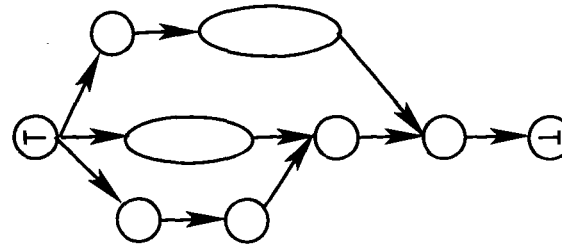


FIGURE 2 : Le treillis est sans cycle et contient toujours un nœud initial et un nœud terminal uniques.

On peut représenter un treillis dans un espace à trois dimensions :

- la première dimension représente le temps (écoulement du texte).
  Appelons T(n) la projection du nœud n sur l'axe des temps. Pour tout arc du treillis, on a :
  $$T(n_d) < T(n_a)$$
  ou n_d est le nœud de départ de l'arc et n_a est le nœud d'arrivée de l'arc ;
- la seconde dimension correspond au niveau d'ambiguïté ;
- la troisième dimension correspond au niveau d'analyse.

Les nœuds de ce treillis peuvent être regroupés en niveaux (on parlera de *strates*) qui correspondent aux niveaux d'analyse. Les nœuds d'une strate peuvent être de natures différentes, mais il est intéressant pour une architecture particulière de restreindre une strate à un seul type de nœuds.

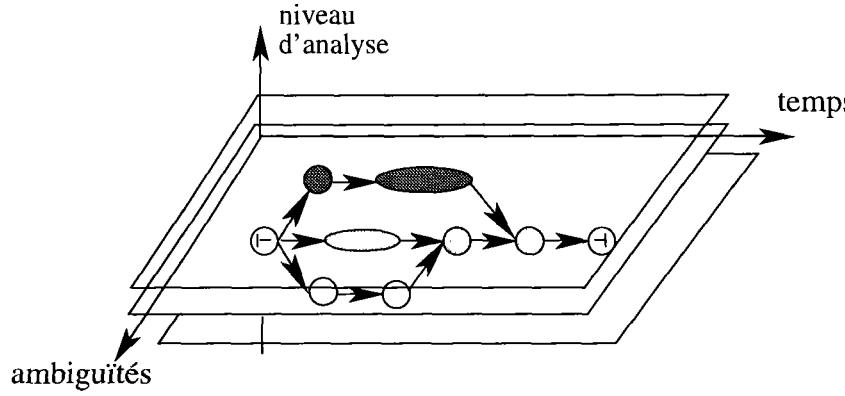
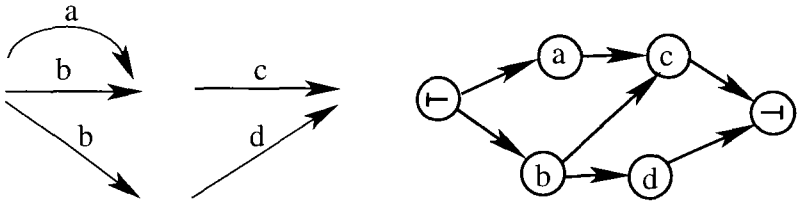


FIGURE 3 : Le treillis se décompose en strates, les nœuds d'une strate dominent les nœuds de la strate inférieure.

L'approche multistrates reste générique et permet de résoudre certains problèmes de TALN. Chaque strate permet de regrouper des nœuds contenant des informations homogènes. Ainsi un outil qui ne manipule qu'un type d'informations, ne trouvera sur une strate que des informations pertinentes. Toutefois le modèle n'interdit pas la définition d'architectures monocouches (avec des nœuds hétérogènes).

Deux nœuds du treillis sont reliés au plus par un arc. Contrairement aux Charts et aux Systèmes-Q (Colmerauer 1970) les informations sont portées par les nœuds. Les arcs ne contiennent qu'une valuation. Deux nœuds ne peuvent pas partager une même structure.

Comme avec les Systèmes-Q, un treillis permet de représenter des alternatives (Boitet 1988), ce qui n'est pas possible avec des Charts. D'autre part, si on essaye, par exemple, de représenter les alternatives {ac, bc, bd}, il est nécessaire de dupliquer l'étiquette d'un arc avec les Systèmes-Q (l'étiquette b) :



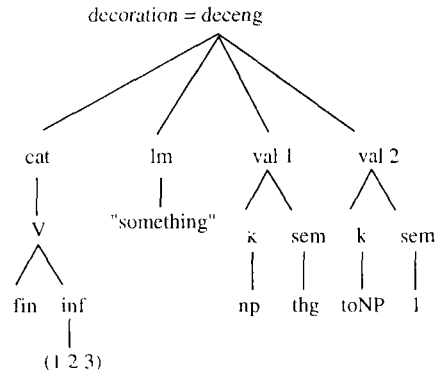
Systèmes-Q Treillis

Décorations

Chaque nœud du treillis peut porter des décorations. Le modèle LEAF n'impose pas de restriction sur le type de décoration utilisé. Il est toutefois intéressant de proposer un type de décoration et une notation standard.

Une décoration est un arbre fini avec une valeur sur chaque nœud. Deux nœuds frères ne peuvent porter la même valeur. Une valeur est une instance de type simple (symbole, nombre, booléen, chaîne de caractères) ou une instance de type complexe (décoration, liste, ensemble, vecteurs,...). La valeur *null* est une instance valide de tous les types dénote la valeur nulle.

Par exemple, la décoration suivante : se notera :




```
deco =
  deceng.[
            cat.V.[fin inf.(1 2 3)]
            lm.«something»
            val1.[k.np sem.thg]
            val2.[k.toNP sem.1]
          ]
```

On note la valeur d'une décoration sous forme de couples attribut-valeur séparés par un «.». Les vecteurs sont notés entre crochets, les listes entre parenthèses.

Par exemple `cat = V.[fin inf].` se lira « la décoration `cat` prend pour valeur `V.[fin inf]`» et `cat.V = [fin inf]` se lira comme « l'attribut `V` de `cat` prend pour valeur `[fin inf]`».

Moteurs

Les moteurs agissent sur la structure. Ils forment la composante active du modèle LEAF. Les moteurs peuvent être conformes au modèle LEAF selon deux niveaux :

conformité limitée : l'interface entre le moteur et la structure se fait via une traduction. Le moteur utilise sa propre structure de donnée ;

conformité complète : le moteur agit directement sur la structure.

La seconde approche peut parfois être difficile à mettre en œuvre lorsqu'un moteur est basé sur un formalisme éloigné de la structure du modèle LEAF. Par contre cette approche a le mérite de fournir un environnement intégré où des moteurs partagent la structure de données. Elle permet ainsi de tirer avantage d'un environnement de programmation linguistique homogène (qui simplifie la mise au point en offrant des outils standard de trace, d'évaluation partielle, de visualisation,...).

Le GETA étudie actuellement l'adaptation des LSPL ATEF (Chauché 1975) et ROBRA (DSE1 1982) au modèle LEAF.

Structures générales de nœud et d'arc

On définira, dans cette section, la structure générale des nœuds et des arcs du treillis. Cette description se fait en termes de classes (au sens usuel des langages à objets).

Une architecture basée sur le modèle LEAF héritera de ces structures et les spécialisera selon ses besoins.

La définition des classes génériques de nœuds et d'arcs se fait à l'aide de la notation exposée ci-dessus. Le premier argument s'interprète comme la liste des classes dont hérite la classe. Les arguments suivants consistent en une suite de paires attribut/type.

Nœud

Un nœud du treillis a la structure suivante :

```
generic-node =  [()
      label.    object
start-bound.    number
  end-bound.    number
        val.    number]
```

- le premier argument « () » indique que generic-node n'hérite d'aucune classe ;
- l'attribut `label` dénote l'étiquette du nœud ;
- les attributs `start-bound` et `end-bound` permettent de définir une mesure de la couverture du nœud selon l'axe du temps ;
- `val` dénote la valuation de l'arc.

Arc

Un arc du treillis a la structure suivante :

```
generic-arc = =  [()
   val.      number]
```

Dans certains contextes, il pourrait s'avérer intéressant d'avoir plus d'une valuation portée par les arcs. Par exemple, plusieurs grammaires décrivant des styles d'énoncés différents pourraient être utilisées (en analyse et en génération) et produiraient des solutions avec des valuations différentes sur les arcs (et/ou les nœuds).

Processus

Plusieurs stratégies sont possibles pour construire ou modifier une strate.

Approche « classique »

Une application linguistique construit une nouvelle strate à partir de la dernière strate calculée. Toutes les analyses sont produites au cours d'un seul processus. Il s'agit d'une énumération exhaustive de l'espace de recherche des solutions.

Approche « semi-classique »

Il s'agit de pouvoir reconnaître les ambiguïtés et de les traiter directement. Ici, les règles de traitement de l'ambiguïté sont confondues avec les règles de construction.

Programmation ambiguë

Les règles de traitement de l'ambiguïté sont ici séparées des règles de construction. Ces règles sont attachées aux cas d'ambiguïtés. Elles peuvent être soit précalculées (il s'agit de schémas) soit elles peuvent être recherchées par le système (détection de plusieurs solutions sans *reconnaissance* du type d'ambiguïté).

L'utilisation des pondérations permet de spécifier des heuristiques portant sur le choix des chemins à traiter en priorité.

Idée d'heuristique

Il s'agit de ne calculer qu'une fraction de l'espace de recherche afin d'éviter toute explosion combinatoire. À nouveau, les règles d'heuristique doivent être définies séparément des règles de construction et des schémas d'ambiguïtés. Ces règles d'heuristique ne doivent pas être figées mais être paramétrables.

Les règles d'heuristiques sont parfois définies à l'extérieur des grammaires mais sont dans ce cas figées (c'est le cas de Prolog avec le mécanisme de retour-arrière).

La recherche se fait donc en programmation dynamique avec une évaluation paresseuse où les solutions sont calculées à la demande. Si les traitements ultérieurs invalident la dernière solution proposée, le système calcule la suivante (s'il en existe).

Autres traitements

Dans le cas de la TA et une approche analyse-transfert-génération, on aura une structure symétrique entre analyse et génération. La dernière strate devra représenter le texte en langue cible.

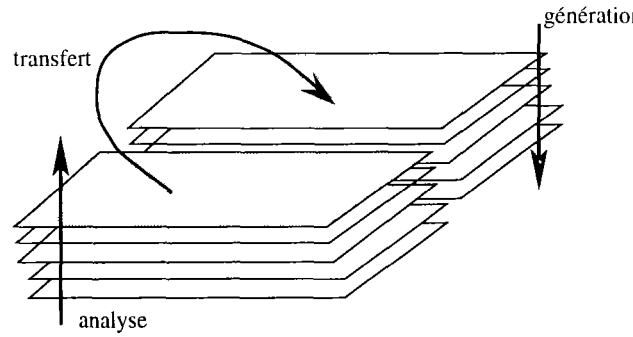


FIGURE 4 : En TA, la structure de treillis est dédoublée et symétrique. Les strates de même nature correspondant à la langue source et à la langue cible sont placées au même niveau sur la figure.

Factorisation des nœuds

Plusieurs nœuds avec des décorations équivalentes peuvent être créés par les moteurs.

Dans certains cas, ces nœuds sont factorisables. Faut-il dupliquer le nœud et associer à chaque copie du nœud une décoration ou bien ne créer qu'un nœud portant comme information une liste non bornée de valeurs de décorations ? L'importance attachée à la géométrie peut être à l'origine d'une stratégie. Factoriser le résultat est plus économique mais il est (parfois) indispensable de dupliquer le nœud en fonction de ce qui va « pointer » au-dessus. En factorisant, on perdrait les liens verticaux (c'est-à-dire les arbres). Une gestion automatique de ce genre de problème est envisageable.

Un exemple d'application : l'interprétation de textes écrits

Cette partie présente un exemple d'architecture basée sur le modèle LEAF, pour l'interprétation des textes écrits. Cette architecture met en jeu des strates simples et des strates complexes définissant chacune des types de nœuds particuliers. La structure des arcs n'est pas changée.

Les strates simples contiennent les lettres pour la *strate de caractères* et les mots (ou morphes) pour les *strates morphologiques*. Les *strates linguistiques* permettent de s'éloigner de la structure de surface et de s'approcher d'une structure d'interprétation.

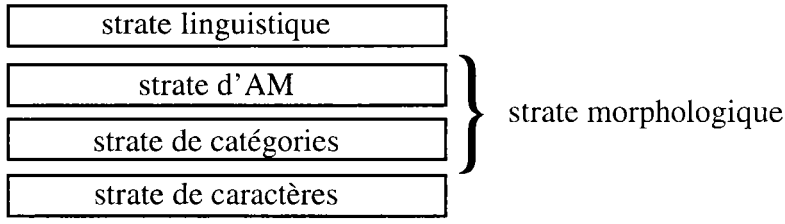


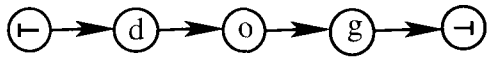
FIGURE 5 : L'architecture LEAF pour l'interprétation des textes écrits définit quatre strates. La strate linguistique peut éventuellement être dépliée en d'autres strates (voir la section « Tranches imbriquées et déploiement »).

Strate de caractères

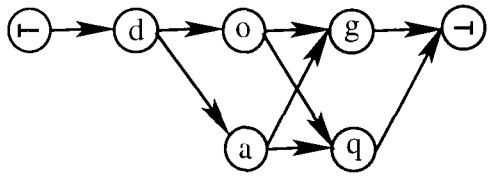
La strate de caractères est la représentation du texte d'entrée sous forme de treillis. Selon la forme et la nature du texte d'entrée, la géométrie du treillis peut varier, cependant l'information portée par l'étiquette du nœud sera toujours un caractère (une chaîne de caractères de longueur unitaire).

```
character-node  =  [(generic-node)
        label.      chaîne]
```

Si le texte d'entrée a été saisi directement sous un traitement de texte (le cas le plus favorable) alors le treillis sera réduit à une chaîne de nœuds contenant les caractères. Il n'y a pas d'ambiguïté dans ce cas. Par exemple, pour le mot *dog* on aura la treillis suivant :

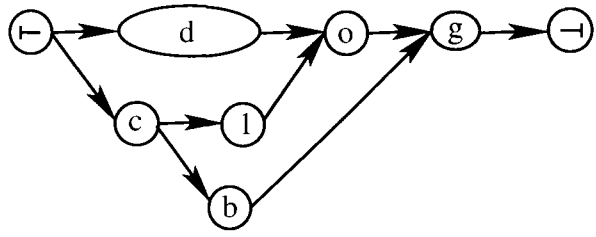


Si le texte vient d'un outil de reconnaissance optique de caractères (OCR), les caractères peuvent déjà présenter des ambiguïtés, on aurait alors :



On a, ici, simplifié les cas potentiels d'ambiguïtés, avec par exemple le *o* reconnu comme un *0* (zéro) ou le *g* comme un *9*.

Dans le cas d'un système de reconnaissance de l'écriture manuscrite, l'entrée pourrait être encore plus ambiguë. La segmentation des lettres de dog poserait alors quelques problèmes :



Dans ce cas, on voit que la couverture (dans le temps) de chaque nœud du treillis n'est pas constante, les nœuds *d* couvrant autant que les nœuds *c* et *l*. On a supposé pour notre exemple que ce genre de problème ne survenait pas avec un système d'OCR.

Strates morphologiques

Il est nécessaire de représenter le texte d'entrée sous forme de « mots », ce qui implique une phase de segmentation. L'idée la plus triviale consiste à regrouper les caractères entre les espaces parmi les nœuds du niveau inférieur. Une correction ty-

pographique est d'ailleurs possible à l'occasion (suppression des espaces en trop) ou une segmentation plus élaborée (à l'aide de lexiques non structurés) afin de gérer les espaces manquants.

Dans le cas de langues posant des problèmes de segmentation (comme le thaï), un tel processus, consistant à construire la strate de mots, serait nécessaire.

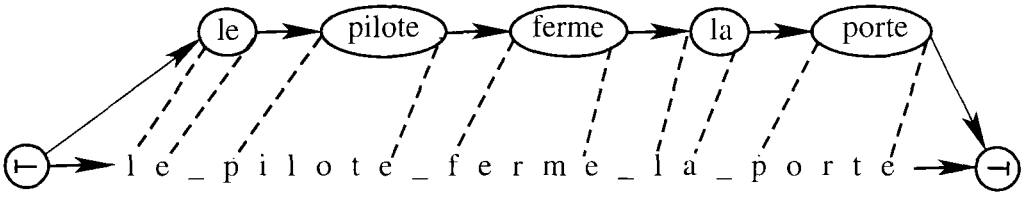


FIGURE 6 : La strate de mots est construite sur la strate de caractères. Cette strate peut déjà être ambiguë pour les langues posant des problèmes de segmentation.

On va considérer deux strates morphologiques : la première est une *strate de catégories,* la seconde est une *strate d'AM* (analyse morphologique). La motivation est que la catégorisation peut se faire avec des processus légers et rapides et peuvent donc « débroussailler à moindre coût » l'analyse morphologique.

Strate de catégories

Chaque nœud de la strate de catégories contient une étiquette et une décoration « légère ». L'étiquette d'un nœud correspond à sa couverture de la strate de caractères, c'est-à-dire à la concaténation des étiquettes des nœuds de caractères qu'il couvre.

La structure de décoration décrit, au plus, le résultat d'une catégorisation, qui peut suivre le schéma suivant :

```
          tagged-node =   [(generic-node)
                decor.    tagged-decoration]
tagged-decoration =   [(generic-decoration)
                  cat.    (verb adjunct pronoun noun
                          deictor subordinator coordinator unknown)]
```

Les catégories ne sont proposées qu'à titre indicatif.

Le calcul de ces catégories peut se faire par des moyens simples et efficaces et peut grandement accélérer des traitements plus complexes comme une analyse morphologique ou syntaxique.

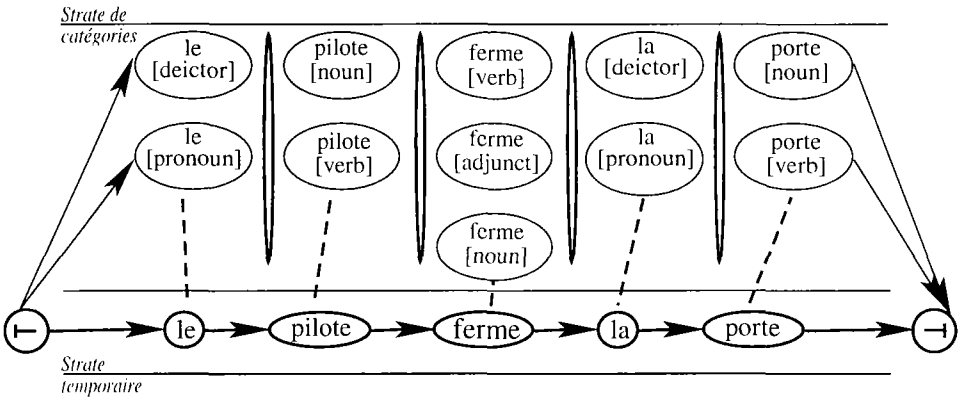


FIGURE 7 : Afin de faciliter l'écriture et la lecture des treillis, on a remplacé les arcs par des losanges verticaux quand chacun des nœuds se terminant à une borne i était lié à tous les nœuds débutant à la borne i+1.

La construction de la strate de catégories se fera en deux étapes :

- création d'une strate de mots (les catégories sont vides) constituée d'une chaîne de nœuds (s'il n'y pas d'ambiguïté) dont les étiquettes constituent le résultat de la segmentation ;
- catégorisation de chacun des nœuds et duplication en cas de catégorisations multiples.

Sur les figures 6 et 7, le treillis de mots n'est en fait que temporaire car il est modifié de façon destructive par le traitement. Une fois la strate de mots construite, la catégorisation viendra dans un deuxième temps décorer les nœuds. On pourrait toutefois considérer comme souhaitable de conserver telle quelle cette strate de mots si on s'intéresse particulièrement au problème de la segmentation.

Strate morphologique

La seconde strate, la *strate morphologique,* correspond au résultat d'une analyse morphologique dans un système de TA (par exemple, une analyse avec ATEF, Chauché 1975). Une telle analyse, calculant des informations exprimées par des morphèmes grammaticaux (comme le genre, le nombre, le mode, le temps, la personne,...) nécessite l'utilisation relativement lourde d'un LSPL.

Selon les langues et leur degré de « morphémisation » on sera à même de représenter plus ou moins d'informations.

On définit la classe de nœud morphologique suivante :

```
morphological-node =   [(generic-node)
                 decor.    morphological-decoration]
```

Une syntaxe utilisable par le linguiste permet d'écrire les définitions de types et

décorations linguistiques (ces objets pourront, au choix du linguiste, être également édités dans des formulaires) :

```
type morphological-decoration
    = all (
          verb : verb-d,
          adjunct : adjunct-d,
          pronoun : pronoun-d,
          noun : noun-d = exc ( nombre : nex (
                                        sing : booléen,
                                        plur : booléen),
                                genre : genre-d,
                                deictor : deictor-d,
                                subordinator = subordinator-d)

decor decor1 : genre-d = exc (
                              masc : booléen,
                              fem : booléen,
                              neut : booléen)
```

Le mot-clé `all` indique que tous les attributs sont obligatoirement présents (une décoration d'un type non exclusif aura donc comme valeur d'initialisation tous ses attributs présents avec comme valeur `null`). Le mot-clé `nex` indique que les attributs sont tous potentiellement présents. Une décoration de ce type peut donc avoir comme valeur n'importe quelle partition sur l'ensemble des attributs. Le mot-clé `exc` indique que la décoration est exclusive, elle ne peut donc instancier qu'un seul de ces attributs à la fois.

On remarquera que la syntaxe permet d'imbriquer ou de séparer les définitions des décorations et de leurs types ainsi que de décrire des types anonymes.

Si on transcrit les définitions écrites par le linguiste en termes de classes liées à l'implémentation actuelle, on obtient les définitions suivantes (on utilise ici la syntaxe définie lors de la présentation du modèle LEAF) :

```
morphological-decoration =     [() all
                        verb.    verb-d
                      adjunct.   adjunct-d
                      pronoun.   pronoun-d
                         noun.   noun-d
                      deictor.   deictor-d
                 subordinator.   subordinator-d
                  coordinator.   coordinator-d
                      unknown.   booléen)]

gender-d =        [() exc          noun-d =    [() exc
     masc.        booléen            nombre.   nombre-d
      fem.        booléen             genre.   gender-d
     neut.        booléen]           commun.   booléen
   anon1 =        [() nex            propre.   booléen]
     sing.        booléen
     plur.        booléen]
```

`decor` est une instance de la classe `genre-d` et pourra prendre, par exemple, la valeur suivante :

```
decor = masc.vrai.
```

On n'en détaillera pas plus avant la description des variables morphologiques.

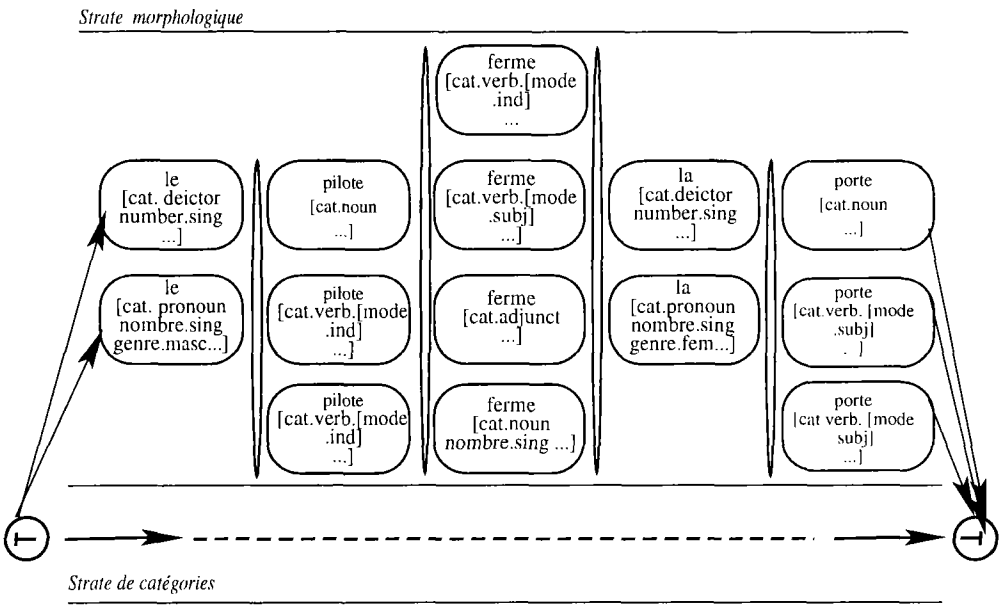


FIGURE 8 : La strate morphologique comporte des nœuds de type morphological-node couvrant la strate de catégories.

Implémentation d'outils

Une implémentation en CLOS (*Common Lisp Object System*) d'ATEF a été réalisée au GETA permettant la construction de cette strate morphologique. L'application du modèle LEAF s'est faite ici dans un contexte de TAO. Le passage de la strate de caractères à la strate de mots n'a pas posé de problèmes spécifiques car les langues sources ne présentent pas de difficultés de segmentation. Les performances satisfaisantes d'ATEF nous ont permis de passer directement de la strate de mots à la strate morphologique (sans passer par la strate de catégories).

Strates linguistiques

L'idée de base est que plus l'analyse monte de niveau (ou que l'utilisateur se déplace profondément dans la structure des décorations) plus on s'éloigne de l'expression de « surface ». Ce qui était implicite (et donc non représenté) devient explicite (représenté).

Nœud linguistique et arbres concrets

Chaque nœud des niveaux linguistiques contient la racine d'un *arbre concret* (arbre de dérivation). Dans le cas d'ambiguïté, le même nœud peut être la racine de plusieurs arbres concrets (voir la discussion sur la factorisation des nœuds). Chaque nœud portera également des *arbres abstraits* décorés correspondant aux arbres concrets (voir la prochaine section).

Un nœud de la couche linguistique a la structure suivante :

```
linguistic-node ==    [(morphological-node)
            d-tree.     concrete-tree-ds
            r-tree.     decorated-abstract-tree-ds]
```

La frontière des arbres concrets (image des feuilles) est en correspondance directe avec le treillis des caractères (c'est-à-dire le texte d'entrée).

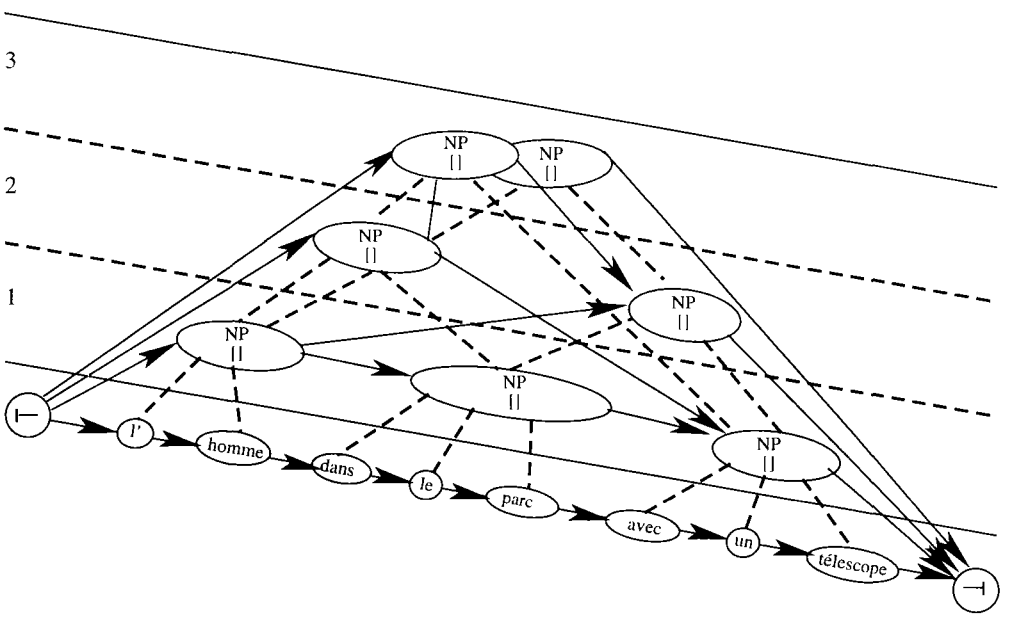


FIGURE 9 : « L'homme dans le parc avec un télescope » a deux analyses possibles selon le rattachement des groupes simples. Les nœuds des strates inférieures 1 et 2 sont factorisés.

Il existe des algorithmes efficaces (par exemple Quinton 1980 et Tomita 1986) qui calculent les arbres concrets pour le treillis de caractères en fournissant un symbole non terminal (ou l'axiome de la grammaire) dans la racine de l'arbre. Les ambiguïtés peuvent alors être factorisées et gérées implicitement ou explicitement.

On ne détaillera pas ici la structure des décorations des nœuds linguistiques.

Visualisation

Une première maquette de visualisation d'un processus de construction de treillis a été réalisée sur HyperCard (sur Macintosh). Il a été vérifié que l'approche consistant à conserver les différents niveaux d'analyse était mieux perçue par le linguiste que l'approche destructive (simplement pour l'appréciation des processus mis en jeu).

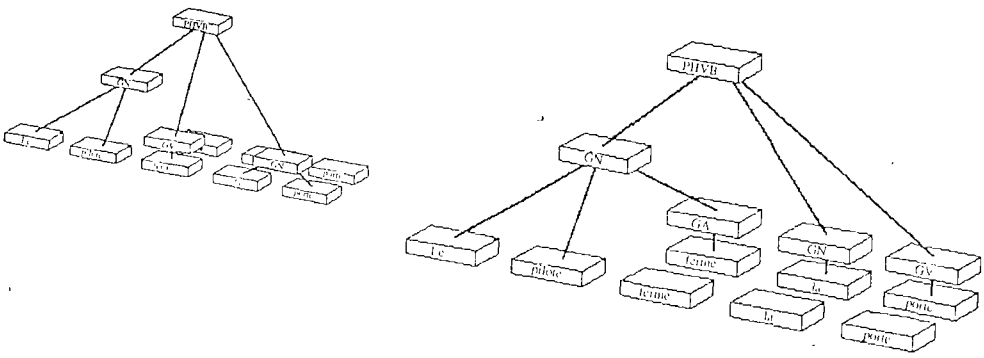


FIGURE 10 : Résultat simulé d'analyses syntaxiques avec deux interprétations possibles. Le linguiste peut mettre en évidence chacun des arbres concrets. Certains nœuds du treillis sont factorisés (ils portent des décorations simples ayant les mêmes valeurs).

L'importance de la visualisation de structures complexes a été particulièrement mise en évidence dans Quinton (1980). L'approche « en volume » de la structure de treillis est perçue comme plus naturelle que des projections n'offrant à chaque instant qu'une perception réduite du résultat d'analyse. Manipuler directement le treillis (le faire tourner, masquer certaines parties, etc.) permet d'avoir une perception immédiate du résultat dans sa globalité.

D'autre part, garder les niveaux précédents peut servir de « filet de sécurité » en cas d'échec des analyses ultérieures, permettant de mettre en place plus efficacement des « règles de formes inconnues ».

Tranches imbriquées et déploiement

Si l'on veut représenter de l'information d'ordre interprétatif (qui sort du cadre strictement linguistique, mais est lié au discours ou au domaine), il est nécessaire d'avoir une structure relativement économique pour porter les décorations (structure d'interprétation). Les arbres abstraits sont bien plus économiques que les arbres concrets ; cependant ils n'ont pas de caractère projectif permettant de lire la chaîne d'entrée sur les mots des feuilles.

Niveaux imbriqués

Il serait intéressant d'ajouter des arbres abstraits comme informations attachées aux nœuds des arbres concrets du niveau linguistique. L'information portée par les nœuds des arbres abstraits est représentée par des psi-termes (Aït-Kaçi 1986).

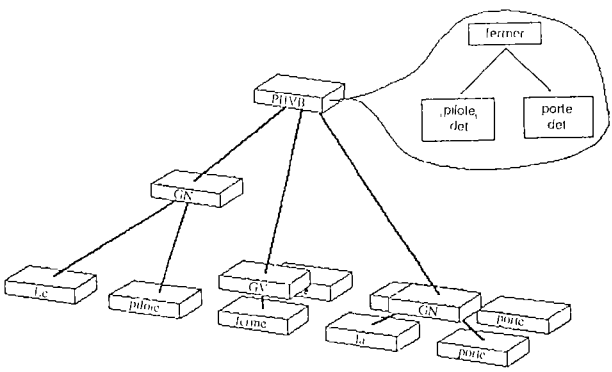
La différence entre les structures simples présentées ci-dessus et les psi-termes réside principalement dans l'absence de réentrance, chez ces dernières. Le niveau de récursivité est donc plus élevé avec les psi-termes. De plus, les psi-termes sont assimilables à des prototypes (au sens des langages à objets) et constituent la seule classe d'objet avec laquelle le linguiste peut travailler. Les structures présentées ici mettent l'accent sur la définition de types et de décorations et sont donc assimilables aux langages de classes et instances.

Arbres abstraits

Les arbres abstraits décorés contenus dans les nœuds des niveaux linguistiques sont imbriqués dans chaque nœud contrairement aux arbres concrets dont les nœuds sont représentés dans le treillis.

```
abstract-tree-node = =      [()
                     decor.     psi-term]
```

L'arbre abstrait représentant tout le sous-arbre concret est porté par chaque nœud. Une factorisation des nœuds est donc ici souhaitable, puisque les sous-arbres des arbres abstraits seront aussi contenus dans les nœuds des sous-arbres concrets.



Ces informations sont relativement coûteuses à calculer ; il serait donc judicieux de ne les générer que sur les meilleurs arbres concrets en présence. Chaque psi-terme contenu dans les nœuds de l'arbre abstrait est calculé en fonction de ces nœuds fils mais aussi des informations linguistiques et des valuations et (peut-être) d'informations sur le discours ou le domaine.

Les raisons sous-jacentes à l'utilisation d'arbres abstraits (plutôt que de placer directement ces informations sur les nœuds des arbres concrets) sont les suivantes :

● ces informations doivent être séparées des résultats linguistiques afin d'alléger la structure des nœuds concrets, d'accroître la modularité et l'adaptabilité du système ;
● ces informations sont lourdes à gérer, les arbres abstraits sont beaucoup plus économiques en nœuds que les arbres concrets. Cela se paye par la non-projectivité des arbres abstraits sur le texte d'entrée (on peut toujours la recalculer si nécessaire).

Déploiement

Il est bon de pouvoir sortir certaines informations incluses dans les nœuds afin d'en faire au choix :

● de nouveaux nœuds dans une même strate ;
● de nouvelles strates.

Avec une représentation directe des informations habituellement contenues dans les nœuds, l'interaction peut être grandement facilitée. Par exemple, le linguiste pourrait désirer voir le treillis se construire en temps réel et vérifier automatiquement certaines informations. Il ne s'agit pas ici de calculer de nouvelles informations, mais de créer une représentation nouvelle de données déjà calculées (et non directement accessibles).

Il est possible de déplier un niveau imbriqué pour en faire une nouvelle strate. Par exemple, les arbres abstraits contenus dans les nœuds du niveau linguistique peuvent « être sortis » de ces nœuds. La nouvelle strate ainsi construite, viendrait se placer au-dessus de la strate linguistique.

La tranche supérieure, ainsi nouvellement créée dispose de nœuds autonomes. Le linguiste préférera travailler sur les structures d'arbres abstraits pour le transfert en TAO. Il est plus pratique pour lui d'en faire une nouvelle tranche qu'il peut manipuler directement. On peut voir dans ce choix une inspiration des Systèmes-Q où les arcs contiennent également des arbres abstraits manipulés par les linguistes.

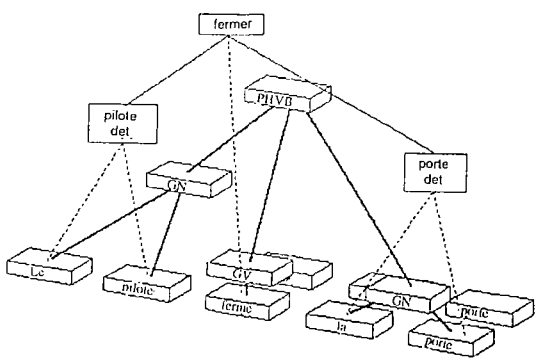


FIGURE 11 : Déploiment de l'arbre abstrait. Un lien entre les nœuds de l'arbre concret est préservé.

La couverture des nœuds des arbres abstraits est souvent discontinue comme dans : *he picks her youngest boy up* où la racine (*to pick somebody up*) couvrira directement les nœuds *picks* et *up*.

Nœuds hétérogènes

Jusqu'à présent, on n'a considéré que des strates dont les nœuds avaient la même structure ; cependant il peut être intéressant de visualiser des informations de nature différente directement dans le treillis.

Nœud de règles

Pour la mise au point, il s'avère particulièrement pratique de garder une trace des règles ayant participé à la construction des nœuds du treillis. On pourra donc faire calculer cette information (au fur et à mesure du traitement) par les outils et la conserver comme une décoration particulière des nœuds.

Cependant, il est plus intéressant d'exhiber ces informations sous forme de nœuds. Le linguiste peut alors veiller au bon déroulement d'un processus et interagir directement avec les nœuds contenant des instances de règles (c'est-à-dire une règle et un état de l'environnement de l'outil au moment de l'application de cette règle).

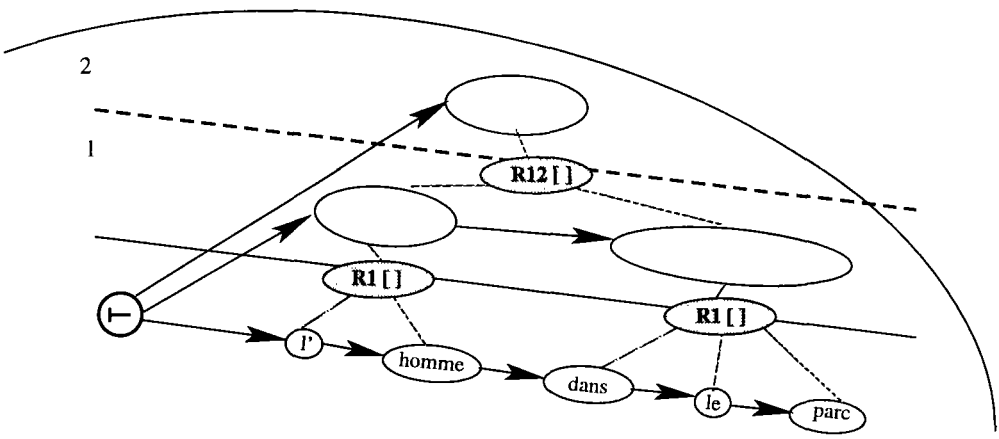


FIGURE 12 : Des nœuds « instances de règles » sont créés directement dans le treillis.

Les instances de règles sont donc accessibles à partir des nœuds du treillis mais aussi à partir des fonctions d'accès aux nœuds du treillis (comme pour les nœuds des arbres concrets).

Une implémentation, en CLOS du modèle LEAF est en cours au GETA. Elle prend la forme d'une collection de classes ainsi que d'un protocole extensible. L'utilisation de Lisp offre une grande souplesse de manipulation pour la programmation dynamique où le linguiste crée indirectement de nouvelles classes de nœuds par dérivation des classes déjà définies.

Conclusion

LEAF est un modèle générique mis au point pour répondre à un certain nombre de problèmes rencontrés en TALN. Le plus crucial de ces problèmes est de permettre une gestion efficace et pratique de l'ambiguïté mais aussi d'offrir une base commune à différents types d'applications linguistiques. En plus de la traduction automatique, on pourra retenir des applications comme la synthèse de la parole à partir de textes (*Text To Speech*), la reconnaissance de la parole (*Speech To Text*), la reconnaissance de l'écriture et d'une manière générale toutes celles faisant intervenir des traitements (linguistiques) complexes et devant gérer l'ambiguïté.

Plusieurs architectures sont dérivables à partir de ce modèle ; celle que nous avons décrite était une solution possible adaptée à l'interprétation des textes écrits. Les différentes strates ont été décrites sans que nous nous attachions pour autant à définir exhaustivement les structures de décoration simples ou celles des psi-termes. L'utilisation de strates dès le niveau des caractères permet de gérer dès que possible les sources de problèmes et ce de manière relativement indépendante de la langue ou de la nature du texte d'entrée. Les strates morphologiques permettent de séparer traitements légers et analyses complètes. Les strates linguistiques offrent plusieurs niveaux d'analyse : un niveau strictement linguistique et un niveau d'interprétation représenté par des arbres abstraits. Ces arbres abstraits peuvent être extraits des nœuds afin d'en faire une nouvelle strate.

Certaines perspectives offertes par ce modèle mériteraient une expérimentation poussée. On pensera notamment à des travaux sur l'effet de la visualisation et de l'interaction directe sur les méthodologies de mise au point de programmes linguistiques utilisés par les linguistes. De même, l'utilisation d'algorithmes de réseaux de neurones sur de grands treillis aboutirait certainement à des résultats intéressants.

Remerciements

Je tiens à remercier tout particulièrement G. Sérasset pour sa relecture patiente et ses critiques avisées et Ch. Boitet pour les nombreuses discussions qui ont amené à l'écriture de cet article.

Références

AÏT-KAÇI, H. (1986) : *An Algebraïc Approach to the Effective Resolution of Type Equations,* vol. 45, pp. 293-351.

BOITET, C. (1988) : *Representation and Computation of Units of Translation for Machine Interpretation of Spoken Texts,* Rap. Comp. & AI, 1988.

CHAUCHÉ, J. (1975) : *Les langages ATEF et CETA,* AJCL, microfiche 17, pp. 21-39.

COLMERAUER, A. (1970) : *Les Systèmes-Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur,* Rap. Université de Montréal, n° 43, septembre 1970.

COLMERAUER, A. (1985) : « Prolog in 10 Figures », *CACM*, vol. 28-12, pp. 1296-1310.

DSE1 (1982) : *DSE1 - Le point sur Ariane-78*, Rap. GETA, Ch. Boitet rédacteur, Contrat ADI/Cap-Sogeti-Innovation/GETA, n° (vol. 1, le logiciel), février 1982.

HAUGENEDER, H. et M. GEHRKE (1986) : « A User Friendly ATN Programming Environment (APE) », *Proceedings of the 11th International Conference on Computational Linguistics*, COLING-86, Bonn, pp. 399-401.

HELM, R., MARRIOTT, K. et M. ODERSKY (1989) : *Building Visual Language Programming, Dictionaries in the Electronic Age : Proceedings of the Fifth Annual Conference of the UW Centre for the New Oxford English Dictionary*, Oxford, Oxford University Press, pp. 105-112.

LENAT, D. B., GUHA, R. V., PITTMAN, K., PRATT, D. et M. SHEPHERD (1990) : « CYC : Toward Programs with Common Sense », *CACM*, vol. 33-8, pp. 30-49.

QUINTON, P. (1980) : *Contribution à la reconnaissance de la parole. Utilisation de méthodes heuristiques pour la reconnaissance de phrases*, Thèse d'État, Université de Rennes.

ROBERTSON, G. G., CARD, S. K. et J. D. MACKINLAY (1993) : « Information Visualization Using 3D Interactive Animation », *CACM*, vol. 36-4, pp. 57-71.

TOMITA, M. (1986) : *Efficient Parsing for Natural Language. A Fast Algorithm for Practical Systems*, 201 p.

USZKOREIT, H. (1989) : « Des faisceaux de traits aux types de données abstraits : nouvelles orientations des représentations et traitement linguistiques », *TA Information*, vol. 1-2, pp. 11-35.

11

La gestion de la terminologie et la traduction automatique

Alan MELBY

Université Brigham Young, Provo, Utah, États-Unis

• *Abstract* •

The underlying assumption of this paper is that both human translation (with various computer tools, including terminology management software) and machine translation are here to stay, so increased cooperation is in order. In particular, the paper proposes that ETIF (Electronic Terminology Interchange Format) could be used to facilitate this cooperation. The master copy of a terminology data base could be stored as an ETIF file. Then software could be written to maintain consistency between the terminological data as it is used by human translators and machine translation systems. The machine translation system lexicons would certainly include information that is not found in the ETIF file, but at least it could be verified that human translators and machine translation systems have access to the same source-language terms and (with a given domain) the same target-language equivalents. The paper describes ETIF (which is an SGML application developed in cooperation with the Text Encoding Initiative and now a work item of ISO Technical Committee 37) and mentions efforts toward compatibility between ETIF and MLEXd (another interchange format).

Dans Melby (1992 : 151-154), nous avons exposé un projet de poste de travail du traducteur à trois niveaux d'assistance et nous avons souligné l'importance de se mettre d'accord sur un format universel d'échange des données terminologiques. À ce moment-là, on venait de commencer à élaborer un nouveau format basé sur SGML. Ce nouveau format, sans nom en 1991 et puis baptisé ETIF, a fait bien des progrès depuis deux ans ; le présent texte donne son état actuel. S'il est facile d'admettre que l'échange de données terminologiques peut bénéficier de la gestion de la terminologie pour la traduction humaine, on peut cependant se demander pourquoi est-ce qu'on mentionne la traduction automatique dans le titre ? Les systèmes de traduction automatique

n'ont-ils pas autant besoin de terminologie que les traducteurs humains ? Il est devenu évident, pratiquement pour tout le monde, que, dans un proche avenir (d'ici l'an 2001), la traduction automatique ne traitera qu'une petite partie (problablement moins de dix pour cent) du texte à traduire, et n'est-ce pas le moment pour deux anciens ennemis (les traducteurs humains et les concepteurs de traduction automatique) de faire la paix ?

Avant de plonger dans les détails de ETIF, nous allons identifier les types de traduction automatique pouvant bénéficier de l'échange de données terminologiques. Cette tâche est simple, puisque tout genre de système de traduction automatique peut en bénéficier sauf un : un système clos dont les dictionnaires n'ont jamais besoin de mise à jour. Du fait que manifestement presque tous les systèmes de traduction automatique ont besoin de mises à jour régulières et que celles-ci touchent principalement les termes techniques (les mots de vocabulaire général restant relativement stables commes formes morphologiques, malgré leur dynamisme sémantique), il est, en plus, élémentaire de constater que le traducteur humain doit prendre connaissance de ces mêmes termes techniques qui font l'objet de la mise à jour du dictionnaire de traduction automatique. Ne pas mettre les mêmes termes (surtout les termes nouveaux) à la disposition des traducteurs humains et, par échange électronique, à la disposition de ceux qui s'occupent de l'entretien des dictionnaires de traduction automatique – ne pas faire cet échange de terminologie – ne peut mener qu'à une des deux conséquences suivantes : soit les recherches terminologiques se feront en double avec les mêmes résultats pour l'humain et pour la machine (ce qui est inutilement dispendieux), soit l'humain et la machine produiront des traductions contradictoires des équivalents de mêmes termes qui se réfèrent aux mêmes concepts (ce qui mène à la confusion et peut devenir plus cher que le travail en double).

Techniquement, on propose une méthodologie simple : les termes sont mis dans un fichier ETIF[1]. Ce fichier sert à : (1) mettre à jour le système de gestion de la terminologie du traducteur humain ; et (2) mettre à jour le dictionnaire du système de traduction automatique. Il est évident qu'un dictionnaire de traduction automatique détiendra des champs qui n'existent pas dans le fichier ETIF, mais on peut facilement créer un logiciel de comparaison qui indique des « trous » dans le dictionnaire de traduction quant il y a un terme + domaine + équivalent qui ne figure pas dans le dictionnaire de traduction automatique. On dit *terme plus domaine* puisqu'un terme peut se traduire de façon différente dans plusieurs domaines. Il suffit d'établir une table de correspondance entre les noms de domaine dans le fichier terminologique et dans les dictionnaires de traduction automatique. Ce logiciel de comparaison pourrait produire une liste de problèmes potentiels, que l'humain peut examiner.

---

1. Pour se renseigner sur ETIF :
- Si on ne connaît pas SGML : van Herwijnen, E. (1991) : *Practical SGML*, Dordrecht, Kluwer ; ou (plus avancé) Goldfarb, C. F. (1990) : *The SGML Handbook*, Oxford, Oxford University Press.
- Si on ne connaît pas TEI : Sperberg-McQueen, C. M. et L. Burnard (dir) (1990) : *Guidelines for the encoding and interchange of machine readable texts*, Draft version 1.1. (version 2 pending May 1994), The Association for Computers and the Humanities (ACH) ; the Association for Computational Linguistics (ACL) ; et the Association for Literary and Linguistic Computing (ALLC), Chicago, Oxford.
- Si on ne connaît pas ETIF : Melby, A. K., Budin, G., et S. E. Wright (1993) : « Terminology Interchange Format ([E]TIF): A Tutorial », *TermNet News*, vol. 40, février.
The proceedings of the ASTM conference STP 1223 (Standardizing and Harmonizing Terminology: Theory and Practice), Philadelphie, octobre 1993.
ISO DIS 12 200 (ETIF) and WD 12 620 (extensive list of data elements), ISO comité technique TC 37.

Mais pour que ce rêve de coopération se réalise, il faut un format d'échange qui soit accepté par de nombreuses personnes. ETIF devient un tel format. ETIF a son origine dans le « *Text Encoding Initiative* » (TEI), qui a comme but la conception de plusieurs DTD (*Document Type Definitions*) pour SGML. SGML est un standard international (ISO 8879), qui s'emploie de plus en plus pour la description de la structure de toutes sortes de textes. Une DTD est une description formelle (selon le « langage » SGML) d'une classe de documents. Voici un exemple d'entrée terminologique en SGML selon la DTD de ETIF :

```
<termEntry id=te84.11>

    <admin type='domain'> plastics </admin>
    <ref type='bibliographic' target='ISO.472-1988'> p. 84 </ref>
    <admin type='creationDate'> 1993.09.30 </admin>
    <ptr type='relatedTerm' target='te04.06'>

    <tig lang=en>
        <term> thermal degradation </term>
        <gram type=pos> n </gram>
        <descrip type='definition'> The entirety of all deleterious chemical modifications of plastic at elevated temperature. </descrip>
        <note> It is essential to report the temperature and other environmental conditions at which the phenomenon is studied. </note>
    </tig>

    <tig lang=fr>
        <term> d&eacute;composition thermique </term>
        <gram type=pos> n </gram>
        <gram type=gen> f </gram>
        <descrip type='definition'> Ensemble de toutes les modifications chimiques nuisibles d'un plastique &agrave;temp&eacute;rature &eacute;lev&eacute;e. </descrip>
        <note> Il est essentiel d'indiquer la temp&eacute;rature et les autres conditions d'environnement dans lesquelles le ph&eacute;nom&egrave;ne est &eacute;tudi&eacute;. </note>
    </tig>

</termEntry>
```

Le début de l'entrée est marqué par la balise <termEntry> et la fin est marquée par la balise </termEntry>. En général, la fin d'un élément est marqué par une balise dont le nom est précédé d'une oblique (/). Il y a trois sections logiques dans cette entrée :

(a) les renseignements généraux qui s'appliquent à l'entrée entière ;
(b) le « tig » anglais (qui commence par <tig lang=en>) ; et
(c) le « tig » français (qui commence par <tig lang=fr>).

Renseignements généraux

Au début de l'entrée, on voit d'abord un élément administratif qui indique le domaine de connaissance (*plastics*) du concept traité dans cette entrée. Ensuite on trouve un pointeur vers une référence bibliographique qui contient les renseignements principaux de cette entrée. L'attribut *target*= est suivi d'une identification (*ISO.472-1988*) qui doit se trouver comme valeur d'un attribut (ID=) ailleurs dans le « document » terminologique. Le reste des renseignements généraux indiquent la date de création de cette entrée et l'identification d'une autre entrée qui est liée à celle-ci.

Le terme anglais, etc.

Un « tig » est un regroupement d'un terme et d'autres éléments qui donnent des renseignements supplémentaires sur le terme et sur le concept traité dans l'entrée. Le tig anglais dans cet exemple inclut le terme *thermal degradation*, un renseignment grammatical, une définition, et une note.

Le terme français, etc.

Le tig français contient les mêmes sortes de renseignements que le tig anglais.

Il faut noter que les éléments (tels que l'élément *définition*) qui occupent plus d'une ligne devraient continuer à la marge gauche, mais on les a inscrits sous la balise de début pour rendre l'exemple plus lisible.

Il est important de signaler que ETIF a été accepté par plusieurs organismes dans le monde de la traductique :

(1) les comités *terminology* et *translation and computers* de l'ATA (American Translators Association) ;
(2) le comité *terminology* de l'ASTM (American Society for Testing and Materials) ;
(3) la FIT (Fédération Internationale des Traducteurs) ;
(4) Infoterm (Agence pour la terminologie, Vienne) ;
(5) LISA (Localisation Industry Standards Association).

ETIF est aussi employé dans le projet *Esprit TWB II* de la CEE.

Finalement, ETIF pourrait bientôt devenir un standard et, en ce moment, il est déjà considéré comme « DIS » (*Draft International Standard*) dans le comité technique TC 37 de l'ISO.

Pour ceux qui connaissent le projet MULTILEX de la CEE, je mentionne MLEXd[2], un autre format d'échange. MLEXd est plus ambitieux que ETIF (qui ne vise que les fichiers terminologiques). MLEXd est conçu comme un format d'échange pour les données lexicographiques, les données terminologiques et les données de lexi-

2. Pour se renseigner sur MLEXd : *MLEXd – Standards for a Multifunctional Lexicon – Final Report –, MULTILEX Project WP9,* CAP Gemini Innovation, Université de Münster, Philips et l'Université du Surrey.

ques pour le traitement automatique des langues naturelles. Loin d'être contradictoires, ces deux formats pourraient se compléter. Si un dictionnaire pour la traduction automatique peut être converti automatiquement en format MLEXd, il serait possible d'écrire un logiciel de comparaison universel entre un fichier ETIF, contenant de nouveaux termes, et un fichier MLEXd. Le format ETIF ne pourrait pas remplacer le format MLEXd, puisque ETIF se limite aux données terminologiques. Et le format MLEXd ne pourrait pas remplacer le format ETIF, puisqu'il est trop riche et flexible pour être accepté par des terminologues. Mais des discussions récentes, entre représentants de ces deux formats, permettent d'espérer une meilleure coopération.

Références

MELBY, A. (1992) : « Pour le traducteur : un poste de travail à trois niveaux d'assistance », Clas, A. et H. Safar (dir), *L'environnement traductionnel. La station de travail du traducteur de l'an 2001*, Actes du Colloque de Mons (Belgique), actualité scientifique, Sillery, AUPELF-UREF, Presses de l'Université du Québec.

12

Fonctions lexicales dans le traitement du langage naturel

Igor MEL'ČUK

Université de Montréal, Montréal, Canada

• *Abstract* •

Lexical Functions [LFs] are designed to describe systematically and exhaustively such set phrases as heavy RAIN ~ heavy LOSSES ~ heavy PRISON TERMS or [to] give a TALK ~ [to] give a LOOK ~ [to] give [N] a status, *etc. Three directions for their theoretical elaboration are sketched: phraseology (collocations), syntax (different types of extraction, etc.), and lexicography (coverage of restricted lexical cooccurrence).*

1. Three aspects of the use of Lexical Functions in various computational applications are discussed:

- *Collocational aspect, or LFs as a tool for lexical choices (e.g.,* Jean M'A DÉTOURNÉ de cette habitude ⇒ John BROKE ME OF this habit, *where* détourner N de *and* break N from *are both values of LF* ***liquOper***$_1$ (habitude, habit)).
- *Communicative aspect, or LFs as a tool for adapting the syntactic structure of the sentence to its communicative orientation ; the Paraphrase System, designed to carry out the appropriate lexical-syntactic transformations for this purpose, is outlined.*
- *Lexico-cohesional aspect, or LFs as a tool for ensuring text cohesion from the viewpoint of the lexical stock used.*

2. The concept of Lexical Function (normal and degenerated) and that of Standard Lexical Function are defined ; Simple Standard Lexical Functions are introduced. Exemples of Simple Standard LFs ***Magn*** *(intensifier) and* ***Oper***$_1$ *(lignt, or operator, verb) are presented and analysed.*

3. A complete list of Simple Standard LFs (paradigmatic and syntagmatic ones) known to date (54) is presented. Each LF is supplied with examples and linguistic comments. The concepts of Complex LFs and Configuration of LFs are introduced.

Remarques d'introduction

Nous avons découvert les Fonctions Lexicales [FL] il y a plus de 30 ans, lors de notre participation à une expédition géologique dans la zone montagneuse semi-désertique, au sud du Kazakhstan – en vue de les utiliser pour la traduction automatique. Nous avons eu l'idée de ce qui est devenu plus tard (avec la collaboration précieuse de A. Žolkovskij) les FL **Magn** et **Oper**$_1$, quand nous étions en train de chercher une méthode simple permettant d'éviter les milliers de tests ennuyeux nécessaires pour permettre à l'ordinateur de trouver les équivalents russes – déterminés par le contexte lexical – de lexèmes anglais comme HEAVY, IMPORTANT, EXTENSIVE, HIGH, etc., d'une part, et [*to*] DO, [*to*] MAKE, [*to*] GIVE, [*to*] GET, etc., de l'autre. Il suffit de prendre quelques exemples au hasard pour mesurer l'étendue de ce problème, d'ailleurs bien connu des traducteurs : HEAVY [*rain*] correspond en russe à SIL'NYJ [*dožd'*], lit. ‹fort›, alors que HEAVY [*losses*] se rend par TJAŽËLYE [*poteri*], lit. ‹lourd›, et HEAVY [*prison terms*] – par DLITEL'NYE [*sroki zaključenija*], lit. ‹de longue durée› ; GIVE [*a talk*] se traduit par ČITAT' [*doklad*], lit. ‹to read›, mais GIVE [*a look*], par BROSIT' [*vzgljad*], lit. ‹to throw›. Les FL devaient permettre d'établir les correspondances nécessaires de façon directe et logique.

Cet objectif des FL est explicitement formulé dans Žolkovskij et Mel'čuk 1967 ; voir aussi une proposition concrète concernant ce sujet dans Kulagina et Mel'čuk 1968 : 301-302.

Mais dès que le premier jeu des FL a été proposé, il s'est avéré que les FL ont par ailleurs un statut théorique fort important en linguistique. *Primo*, elles constituent le « chaînon manquant » de la THÉORIE DE LA PHRASÉOLOGIE, parce qu'elles permettent de décrire de façon rigoureuse et systématique les ***collocations*** – qu'on pourrait appeler également ***semi-phrasèmes***. Cette contribution théorique des FL est traitée, de façon plus détaillée, dans Mel'čuk 1994.

Secundo, elles sont essentielles dans la THÉORIE DE LA SYNTAXE, car plusieurs régularités syntaxiques ont besoin du concept de FL pour se prêter à une description formelle satisfaisante. Voir, par exemple, (1), tiré d'Abeillé 1988 :

(1) anglais

a. *King John launched* [=la FL **IncepOper**$_1$] *an attack against the city.*
VS
Which city did King John launch an attack against?
It is against this city that King John launched an attack.

b. *King John watched an attack against the city.*
VS
**Which city did King John watch an attack against?*
**It is against this city that King John watched an attack.*

Comme on peut le constater, l'extraction dans des constructions de ce type est possible ou impossible selon que le verbe en question est ou n'est pas une FL de son CO^{dir}.

Tertio, les FL ont ouvert des perspectives fort prometteuses dans la THÉORIE DE LA LEXICOGRAPHIE.

Dans cet article, nous n'expliciterons pas toutes les FL et leur apport théorique maintenant et nous nous limiterons à trois sujets : les usages possibles de FL dans le

traitement du langage naturel par ordinateur, la définition du concept même, et l'inventaire des FL (tel qu'il se présente aujourd'hui).

Nous aimerions souligner que cette présentation s'appuie essentiellement sur des notions générales puisées dans la Théorie Sens-Texte et surtout dans des travaux liés au *Dictionnaire explicatif et combinatoire du français contemporain* (= DEC ; Mel'čuk *et al.*, 1984, 1988, 1992). Étant donné l'espace limité à notre disposition, nous nous devons de supposer une familiarité suffisante du lecteur avec le DEC.

Fonctions lexicales dans les applications computationnelles

Nous connaissons trois domaines majeurs d'utilisation des FL dans les descriptions linguistiques orientées vers les applications informatiques. On peut dire que ce sont des tâches où les FL sont plus que simplement utiles ou commodes : elles sont indispensables. Ces domaines sont :

- Les FL comme instrument de choix lexicaux corrects au sein des syntagmes semi-figés (toutes les FL) : aspect COLLOCATIONNEL. Entre autres, ces choix peuvent être exploités, dans le contexte computationnel, pour assurer la variété suffisante du texte de sortie (en vue de le rendre plus élégant, plus « humain »). De telles utilisations des FL sont liées au Système de Paraphrasage (Mel'čuk 1992).

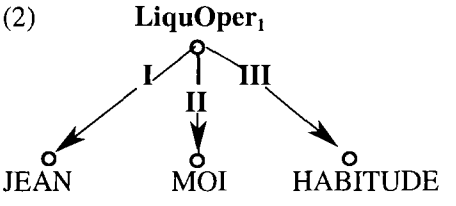
- Les FL comme instrument de choix lexicaux nécessaires pour adapter la Structure Syntaxique Profonde [= SSyntP][1] de la phrase sous synthèse à sa Structure Communicative (certaines FL syntagmatiques verbales : verbes supports, verbes causatifs et verbes de réalisation) : aspect COMMUNICATIF.

- Les FL comme instrument de choix lexicaux nécessaires pour assurer la cohérence du texte sous synthèse (certaines FL paradigmatiques) : aspect LEXICO-COHÉSIONNEL.

Fonctions Lexicales et choix lexicaux collocationnels

L'usage des FL visant à trouver le bon cooccurrent dans une collocation semble assez évident. Ainsi, dans un système de traduction automatique qui fonctionne au niveau de la SSyntP (c'est-à-dire sans passer par la Structure Sémantique), il suffit, dans un premier temps, de réduire la collocation de la langue source à sa représentation au moyen de FL, donc à sa SSyntP ; ensuite on ne traduit que le mot clé de la collocation ; et, finalement, on choisit la valeur de la FL en question pour l'équivalent du mot clé dans la langue cible. Par exemple, la phrase *Jean m'a détourné de cette habitude* est représentée – à l'étape de l'ANALYSE – par une SSyntP comme celle-ci :

(2)

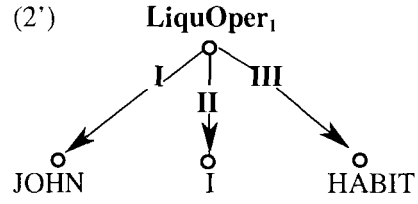


1. Il s'agit, dans cet article, de la syntaxe dite profonde, qu'il nous est impossible d'expliquer ici. Pour cette raison, nous nous limiterons à des références, en nous efforçant d'illustrer les notions pertinentes par des exemples.

Cette analyse s'effectue à l'aide d'un dictionnaire monolingue français (du type DEC) qui présente les valeurs de toutes les FL pour toutes les lexies vedettes ; ainsi, il indique que détourner correspond à **LiquOper$_1$** (HABITUDE).

Ensuite, à l'étape du TRANSFERT, l'arborescence française (2) [= une SSyntP] se voit remplacer par l'arborescence anglaise correspondante (2') :

(2')



Ce transfert s'effectue à l'aide d'un dictionnaire bilingue français-anglais, qui établit les correspondances du type JEAN ⇔ JOHN, MOI ⇔ I, et HABITUDE ⇔ HABIT.

Finalement, à l'étape de la SYNTHÈSE en anglais, la structure (2') se réalise en *John broke me of this habit.*

On remarquera que, dans l'exemple (2), ainsi que dans tous les cas du même type, seuls les noms[2] nécessitent un « vrai » transfert, c'est-à-dire la recherche de leurs équivalents anglais dans un dictionnaire bilingue. Grâce à la méthode proposée, on évite complètement la recherche de correspondances « bizarres » du type DÉTOURNER = [*to*] BREAK dans le contexte de HABITUDE/HABIT : [*to*] BREAK sera calculé comme un élément de la valeur de la FL **LiquOper$_1$**(HABIT), spécifiée sous l'entrée de HABIT dans un dictionnaire monolingue anglais (indépendamment du lexème source DÉTOURNER – ou de tout autre lexème source de n'importe quelle langue). La même entrée fournira le régime : BREAK *whom* ***of*** *what.*

De cette façon, la traduction multilingue des collocations dans toutes les directions voulues n'exige pas plusieurs dictionnaires spéciaux arrangés par paires de langues. Il suffit d'avoir des dictionnaires MONOLINGUES assez détaillés et rigoureux, contenant les valeurs des FL, ainsi que toute l'information pertinente (régime, etc.). Aussi les LF apparaissent-elles comme une INTERLANGUE commode pour le transfert des collocations.

Pour rendre plus claire la procédure de transfert des collocations, je citerai une série de correspondances anglais-français qu'il est facile d'exprimer par des FL (exemple adapté de Fontenelle 1993 ; les FL sont énumérées dans la section relative aux fonctions lexicales standard simples) :

2. Au sens large, y compris les noms propres et les pronoms.

	anglais HABIT ⇔ **français** HABITUDE	
IncepOper$_1$	[to] acquire, develop, form [ART ~], get [into ART ~], take [to ART ~]	contracter, prendre [ART ~]
FinOper$_1$	[to] drop [ART ~], get out, get rid [ART ~], ...	abandonner, perdre [ART ~]
LiquOper$_1$	[to] break [N of ART ~], wean [N from ART ~]	détacher, détourner [N de ART ~]
Liqu$_1$Oper$_1$	[to] break off, kick, shake off, throw off [ART ~]	se débarrasser, se défaire [de ART ~], renoncer [à ART ~], rompre [avec ART ~]
CausFunc$_1$	[to] instill [ART ~ in(to) N]	inculquer [ART ~ à N]

FIGURE 1 : Correspondances collocationnelles exprimées par Fonctions Lexicales.

Ce qui est surtout intéressant dans ce contexte, c'est que les FL sont reliées entre elles par un système d'équations qui fait partie du Système de Paraphrasage déjà mentionné (Mel'čuk 1992). Ces équations assurent elles-mêmes les transformations syntaxiques nécessaires (des SSyntP). Ainsi, la phrase russe *On vzjal zverja na mušku,* littéralement (Il a-pris le-fauve sur guidon [du fusil] = Il a visé le fauve) se réduit, dans l'analyse du russe, à la SSyntP suivante :

III

ON **I**, ← **Labreal$_{12}$**(MUŠKA)**II**,→ ZVER' MUŠKA,

qui, dans le transfert russe-anglais, est remplacée – moyennant le dictionnaire russe-anglais et les règles universelles du Système de Paraphrasage – par la SSyntP anglaise :

III

HE **I**,← **Real$_1$**(BEAD)**II**,→ BEAD BEAST.

Cette SSyntP de l'anglais est aisément réalisée, en synthèse, comme *He drew a bead on the beast.* De cette façon, les FL assurent les ajustements syntaxiques imposés par le transfert des unités lexicales entre des langues données – partout où les FL sont impliquées.

Cependant, si le système de traduction automatique en question passe à travers une RSém, alors la tâche (en ce qui concerne la cooccurrence lexicale restreinte) consiste à déterminer, lors de la synthèse à partir de cette RSém, la FL pertinente et à « calculer » ensuite sa valeur pour la lexie donnée L. Ce calcul s'effectue, comme je l'ai déjà dit, en ayant recours à un dictionnaire monolingue de la langue cible, du type DEC. La même procédure est, bien entendu, nécessaire pour la génération de texte, quelle que soit la représentation sous-jacente aux textes synthétisés.

Un aspect important de l'usage des FL dans le traitement automatique du langage

tient à leur application dans le paraphrasage, qui peut poursuivre deux buts (reliés mais différents). D'une part, les FL peuvent s'avérer indispensables lorsqu'il s'agit de construire une phrase à partir d'une RSém donnée dans des conditions fort contraignantes (lexicales, syntaxiques et morphologiques). Pour aboutir au moins à une phrase, il faut très souvent disposer d'un puissant générateur de variantes alternatives dont au moins une satisfera toutes les contraintes d'usage imposées par la langue. D'autre part, le paraphrasage basé sur les FL est également nécessaire à la production d'un texte assez varié et flexible, libre de monotonie « machinelle » (voir, à ce propos, Iordanskaja *et al.* 1994 et Iordanskaja, dans cet ouvrage). Toutefois, le thème « FL et paraphrasage » représente un sujet vraiment à part, que nous ne traiterons pas ici.

Fonctions Lexicales et Structure Communicative du texte

L'utilisation des FL pour exprimer correctement la Structure Communicative d'une phrase par sa structure lexico-syntaxique est traitée dans Wanner et Bateman 1990. Une présentation détaillée de cette démarche exige une description du Système de paraphrasage (Mel'čuk 1992) et de la Structure Communicative, ce qui est impossible dans le cadre de cette présentation. Par conséquent, je me limiterai à un exemple (adapté de Wanner et Bateman 1990). Supposons que notre système de génération de textes doive produire des phrases exprimant le sens de (3) :

(3) *The adjective « electronic » indicates to the reader that the dictionaries are dedicated to computers.*

Si dans la SSém de (3), le sens du groupe *the adjective « electronic »* est spécifié comme Thème du sens à exprimer, alors c'est la phrase (3) qui sera synthétisée. Cependant, si l'on spécifie comme le Thème de la SSém de départ le sens du groupe *to the reader*, une SSynt différente devient nécessaire, ce qui donnera éventuellement (3') :

(3') *The reader gets an indication that the dictionaries are dedicated to computers from the adjective « electronic ».*

Pour pouvoir remplacer automatiquement *indicate* par *get an indication*, nous avons besoin des deux documents.

Le premier est constitué d'un système d'équations de paraphrasage du type :

$$\mathbf{V \Leftrightarrow S_0(V) + Oper_2(S_0(V))};$$

autrement dit, il s'agit d'équations reliant des FL entre elles. À l'aide de ces équations, s'obtiennent des équivalences comme :

X analyzes Y ⇔ *Y undergoes an analysis by X*
X resists to Y ⇔ *Y runs into a resistance by X*
X orders Y to Z ⇔ *Y receives from X an order to Z,* etc.

Ce système d'équations embrasse toutes les FL et fournit les indications nécessaires quant aux transformations syntaxiques que les remplacements lexicaux suggérés peuvent exiger (pour plus de détails, voir Mel'čuk 1992).

Le deuxième document est un dictionnaire où, pour toute lexie L, les valeurs de toutes les FL applicables à L sont spécifiées. Il s'agit du DEC dont il a déjà été question.

Nous aimerions faire remarquer en passant que la Structure Communicative sert de filtre dans le paraphrasage : ne sont admises que les paraphrases qui ne contredisent pas la Structure Communicative de la phrase à synthétiser (voir Iordanskaja, ce volume).

Fonctions Lexicales et cohésion du texte

Les FL apparaissent également indispensables pour la sélection des expressions référentielles dans des liens anaphoriques – il faut pouvoir varier de telles expressions pour éviter des répétitions fastidieuses, tout en garantissant la cohésion maximale du texte résultant (voir, par exemple, Lee et Evens 1994, Alonso *et al.* 1992 :160-165, Alonso et Tutin 1993). Par exemple, si on parle de *ambush* ‹embuscade›, on peut y référer en décrivant ses participants comme des *attackers* ‹attaquants›.

(4) *An Indonesian* ***patrol*** *was caught in an* ***ambush***. *The* ***attackers*** *fired three rockets at the* ***soldiers*** *and sprayed them with automatic fire* ‹Une patrouille indonésienne a été prise dans une embuscade. Les attaquants ont tiré trois fusées sur les soldats et les ont arrosés de rafales de mitraillette›.

Ici, *attacker* = $\mathbf{S}_1$(*ambush*), and *soldier* = $\mathbf{S}_1$(*patrol*). Ces connaissances lexicales sont utilisées de façon évidente pour construire une séquence cohérente de phrases (4).

Voici un autre exemple :

(5) *Les ventes ont légèrement* ***augmenté*** *au Québec et en Ontario. De modestes* ***gains*** *sont également constatés en Colombie Britannique.*

Au lieu de reprendre la même expression en disant *Les ventes ont légèrement augmenté également en Colombie Britannique,* le locuteur préfère utiliser $\mathbf{S}_2$(*augmenter*) = *gain* [‹le montant par lequel X augmente›] avec son $\mathbf{Func}_0$ (= *être constaté*) ; cela lui permet de produire un texte plus varié et, de ce fait, plus élégant.

Ayant indiqué les avantages des FL, je caractériserai maintenant de plus près leur nature. Je commencerai par la définition des FL en général, pour ensuite définir le type le plus important des FL, à savoir les FL standard.

Le concept de fonction lexicale

Une fonction lexicale [= FL] est une fonction au sens mathématique : une dépendance ou correspondance **f** qui associe à une lexie **L**, appelée l'ARGUMENT de **f**, un ensemble de lexies **f**(L) – la VALEUR de **f**. Chaque FL correspond à un sens très général (qui peut, à la limite, être vide) et à un rôle syntaxique profond ; l'argument d'une FL **f** est la lexie L sur laquelle le sens ‹f› porte ; la valeur de la FL **f** est une sélection de lexies qui peuvent réaliser **f** pour l'argument donné. Plus précisément, pour qu'une telle dépendance **f** soit une fonction lexicale, deux conditions particulières doivent être satisfaites simultanément, ce qui nous amène à la Définition 1.

Définition **1** : Fonction lexicale (FL)

Une dépendance lexicale **f** qui associe à une lexie L d'une langue **L** un ensemble **f**(L) d'expressions lexicales est appelée une ***fonction lexicale*** si et seulement si une des deux conditions suivantes est satisfaite :

A. Ou bien **f** est applicable à plusieurs L_i ; dans ce cas, quelles que soient les lexies L_1 et L_2, si $\mathbf{f}(L_1)$ et $\mathbf{f}(L_2)$ existent toutes les deux, alors :

1. Des éléments quelconques de $\mathbf{f}(L_1)$ et de $\mathbf{f}(L_2)$ entretiennent (à peu près) la même relation avec L_1 et L_2, respectivement, en ce qui concerne le sens et le rôle syntaxique profond :

$$\frac{(L \in \mathbf{f}(L_1))}{(L_1)} = \frac{(L \in \mathbf{f}(L_2))}{(L_2)}$$

2. Au moins, pour certains arguments, $\mathbf{f}(L_1) \neq \mathbf{f}(L_2)$.

B. Ou bien **f** n'est applicable qu'à une seule L (ou peut-être à deux ou trois L sémantiquement apparentées).

Les FL du type A sont appelées des FL ***normales*** ; celles du type B, des FL ***dégénérées***.

Pour les FL normales, la condition 1 caractérise une dépendance lexicale comme une FL POTENTIELLE ; elle ne fait pas appel aux données spécifiques d'une langue particulière. Par contre, la condition 2 caractérise une dépendance lexicale comme une FL ACTUELLE ; elle fait appel aux données spécifiques de la langue **L**. D'un point de vue linguistique, la condition 2 signifie qu'en **L**, les éléments de la valeur de **f** sont PHRASÉOLOGIQUEMENT LIÉS par leur argument.

Quant aux FL dégénérées, elles sont caractérisées indépendamment d'une langue particulière.

Prenons comme exemple la FL **f** = (intensificateur). Illustrons d'abord la condition 1. Soit L_1 = PLEURER et L_2 = PLUIE ; alors :

$\mathbf{f}(L_1)$ = *amèrement, à chaudes larmes, comme une madeleine, toutes les larmes de son corps, comme un veau, comme une vache, comme un enfant ;*
$\mathbf{f}(L_2)$ = *grosse* | prépos, *diluvienne, torrentielle, violente,...*

Tout élément du premier ensemble (par exemple, *comme une madeleine*) se trouve par rapport à PLEURER dans une relation sémantique et syntaxique qui est identique à la relation qu'entretient avec PLUIE tout élément du deuxième ensemble (par exemple, *grosse*) :

$$\frac{\textit{comme une madeleine}}{\text{PLEURER}} = \frac{\textit{grosse}}{\text{PLUIE}} = \ldots$$

Bien entendu, nous ne prétendons pas que *comme une madeleine* et *grosse* sont sémantiquement ou syntaxiquement équivalents. Pourtant, l'expression *comme une*

madeleine remplit par rapport à PLEURER (à peu près)[3] le même rôle que l'adjectif préposé *grosse* par rapport à PLUIE : les deux sont des modificateurs intensificateurs qui signifient dans ce contexte ‹beaucoup›, ‹très›, ‹intense/intensément›. La proportion donnée ci-dessus peut être prolongée *ad libitum.* Pour être une FL, une dépendance lexicale doit donc donner lieu à un grand nombre de proportions de ce genre. Mais bien que nécessaire, cette condition n'est pas suffisante : il faut de plus que la dépendance lexicale en question respecte la condition 2.

L'importance de la condition 2 peut être illustrée comme suit. Si la dépendance **f** considérée donne lieu à des proportions comme celles ci-dessus (donc, si elle satisfait la condition 1) mais que l'on a toujours le même numérateur pour des dénominateurs différents, une telle dépendance **f** est triviale en **L** : elle ne présente pour nous aucun intérêt, puisque le résultat de son application n'est pas une collocation ; nous ne voulons pas que **f** soit qualifiée comme une FL. Ce n'est qu'un simple cas de signification lexicale. Par exemple, le sens ‹cher› [= ‹qui est d'un prix élevé›] ne correspond pas à une FL en français, car avec n'importe quelle lexie il peut toujours être exprimé par le même lexème CHER. Cela signifie que son expression ne dépend pas phraséologiquement de la lexie modifiée :

$$\frac{\textit{cher}}{\text{VOITURE}} = \frac{\textit{cher}}{\text{VOYAGE}} = \ldots$$

Par contre, le sens ‹très› (= ‹intense›) se présente comme une FL en français : ‹très›(*malade*) = *très, gravement* <**grièvement*>, mais ‹très›(*blessé*) = *gravement, grièvement* <**très*> ; ‹très›(*grippe*) = *carabinée,* mais ‹très›(*prix*) = *haut, élevé,...* ; ‹très›(*lutter*) = *sans relâche, à corps perdu,* mais ‹très›(*battre*) = *à plate couture* ; etc. Comme on le voit, l'expression de ‹très› [= de l'idée d'intensité] dépend de la lexie modifiée. Ce sens correspond, en fait, à la FL **Magn** (voir plus loin).

L'argument d'une fonction lexicale (*malade, blessé, grippe, prix,* etc. par rapport à ‹très›) est aussi appelé MOT CLÉ ou LEXÈME CLÉ. Nous sommes obligés d'utiliser cette appellation pour éviter, dans certains contextes, l'homonymie fâcheuse du terme *argument* (d'une FL *vs* d'un prédicat sémantique).

Parmi les FL normales, il convient de distinguer une sous-classe importante, que nous appelons les FL STANDARD. Ces FL satisfont deux conditions supplémentaires, données par la définition 2.

Définition **2** : Fonction lexicale standard

> Une fonction lexicale **f** est une FL ***standard*** si et seulement si les deux conditions suivantes sont satisfaites :
> 3. **f** est définie pour un grand nombre d'arguments. (Autrement dit, **f** a une vaste cooccurrence sémantique : le sens ‹f› est suffisamment abstrait et général pour être compatible avec beaucoup d'autres sens.)
> 4. **f** possède un grand nombre de valeurs différentes. (Autrement dit, l'ensemble de toutes les valeurs de **f** pour tous les arguments est suffisamment grand.)

3. Nous n'entrerons pas dans les détails de la précision sémantique, c'est-à-dire du sens exact de cet « à peu près ».

Parallèlement à la définition **1**, la condition 3 caractérise une fonction lexicale comme une FL standard POTENTIELLE ; elle ne fait pas appel aux données spécifiques d'une langue particulière. Mais la condition 4 caractérise une fonction lexicale comme une FL standard ACTUELLE ; elle fait appel aux données spécifiques de la langue **L**.

Illustrons le rôle de la condition 3 par l'exemple suivant. Le sens ‹sans ajout de produit modifiant le goût› est exprimé de façon fort spéciale avec le nom CAFÉ : ***noir*** ; ainsi, le thé sans lait ni citron ne peut pas être appelé **thé noir* – il faut dire *thé* ***nature*** *<*café nature>*. De même le whisky sans soda etc. s'appelle *whisky* ***sec***. Les expressions de ce sens sont distribuées lexicalement : NOIR avec CAFÉ, NATURE avec THÉ, SEC avec les boissons alcoolisées. De ce fait, le sens ‹sans ajout de produit modifiant le goût› satisfait les conditions 1 et 2 de la définition 1 : il correspond à une FL. Mais il contredit la condition 3 de la définition 2 : ce sens est trop spécifique, il n'est applicable qu'aux noms de boissons. (Il va également à l'encontre de la condition 4 de la définition 2.) C'est une FL ***non standard***.

Pour montrer le rôle de la condition 4, nous citerons un exemple russe. En russe, le sens ‹de couleur brune› a cinq expressions différentes en fonction de ce qu'il caractérise : si on l'applique à un objet quelconque différent des yeux humains, des cheveux humains et de la peau des chevaux, ‹brun› est KORIČNEVYJ. Mais pour dire ‹brun› en parlant des yeux, on a KARIJ : ‹des yeux bruns› = *karie glaza <*koričnevye glaza>* ; pour les cheveux c'est TËMNORUSYJ ou KAŠTANOVYJ (selon la nuance) : ‹des cheveux bruns› = *tëmnorusye* ou *kaštanovye volosy <*koričnevye volosy>* ; enfin, pour les chevaux, on utilise GNEDOJ : ‹un cheval brun› = *gnedoj kon'/gnedaja lošad' <*koričnevyj kon', *koričnevaja lošad'>* (plus précisément, GNEDOJ s'applique si le cheval a une crinière et une queue noires). Par conséquent, le sens ‹de couleur brune› détermine en russe une dépendance lexicale qui satisfait les conditions 1 et 2 de la définition 2 : c'est une FL. De plus, contrairement au sens ‹sans ajout d'un produit modifiant le goût›, le sens ‹de couleur brune› satisfait aussi la condition 3 de la définition 2 : le nombre de choses qui peuvent être brunes est très élevé. Cependant, ce sens contredit la condition 4 : il n'a que cinq expressions différentes, dont quatre (KARIJ, TËMNORUSYJ, KAŠTANOVYJ, et GNEDOJ) sont utilisées chacune avec très peu d'arguments qui, de plus, sont extrêmement spécifiques. Ce sens correspond également à une fonction lexicale non standard.

Les FL non standard ne se prêtent pas à une organisation hiérarchique et systématique. Elles sont nombreuses (probablement des dizaines de milliers dans chaque langue), mais capricieuses et imprévisibles, de sorte que le lexicographe est obligé de les chercher empiriquement pour les consigner dans les entrées lexicales correspondantes : *vin rouge*[4] *<blanc, rosé> ; café crème <au lait, arrosé, irlandais, liégeois,...> ; nuit blanche, ceinture de sécurité, payer comptant, payer rubis sur l'ongle,...* La seule consolation pour le lexicographe est que les FL non standard sont d'habitude fort spécialisées, ont des sens très précis et ne concernent chacune qu'un domaine lexical très particulier. Dans le présent article, nous ne considérerons que les FL ***standard***.

Parmi les FL standard, nous avons établi empiriquement un sous-ensemble d'à peu près SOIXANTE FL qui s'est révélé particulièrement commode pour la description de la cooccurrence lexicale restreinte et du paraphrasage. Chacune de ces FL est identifiée

4. Par exemple, on notera qu'en espagnol, vin rouge se dit *vino tinto <*rojo>*, et en géorgien, *savi gvino*, lit. (vin noir).

par un nom conventionnel et est traitée comme unité ultime, c'est-à-dire indécomposable. Ces FL constituent le noyau du système des FL et se nomment « FL ***standard simples*** ».

Toutes les autres FL standard entrent dans la sous-classe des FL ***standard complexes***. Elles sont construites à partir des FL standard simples, selon quelques règles générales. Dans ce qui suit, nous allons nous concentrer sur les FL standard simples, en nous limitant à quelques illustrations des FL complexes.

Voici deux exemples de FL standard simples, écrites sous la forme adoptée pour la présentation des FL dans la Théorie Sens-Texte.

– La FL **Magn** (les noms de FL viennent toujours du latin ; dans ce cas, de *magnus* ‹grand›), qui est un intensificateur :

Magn(*riposte*) = *foudroyante*
Magn(*cri*) = *d'orfraie* | avec *pousser*
Magn(*applaudissements*) = *nourris, frénétiques*
Magn(*sourd*) = *comme un pot*
Magn(*fort*) = *comme un Turc*
Magn(*saoûl*) = *comme un Polonais, comme une grive <une barrique, une bourrique>*
Magn(*apprécier*) = *grandement*
Magn(*ivre*) = *-mort*
Magn(*recourir* [*à* N]) = *largement*
Magn(*dormir*) = *profondément, comme un loir, comme une bûche<une souche>, à poings fermés*
Magn(*surveiller*) = *étroitement, de près*

– La FL **Oper**$_1$ (lat. *operari* ‹travailler›), qui est un verbe sémantiquement vide (ou vidé de son sens dans le contexte de son mot clé) et qui prend :

1) le mot clé [= X] comme son complément d'objet direct (CO^{dir}) ou principal (si le verbe n'est pas transitif), c'est-à-dire comme son actant Synt(axique) P(rofond) **II**[5] ;

2) l'actant SyntP **I** potentiel [= Y] du mot clé comme son sujet grammatical (SG), donc comme son actant SyntP **I** :

o Oper$_1$(X)

I **II**

Y **o** **o** X

où Y est l'actant SyntP **I** de X. Voici quelques exemples :

Oper$_1$(*plainte*) = *porter* [~]

(6) *Jean* [= Y] *a porté* [= Oper$_1$] *plainte* [= X] *contre le préposé aux dossiers étudiants,*

où JEAN est l'actant SyntP **I** de PORTER, alors que PLAINTE est son actant SyntP **II** ; en même temps, JEAN est l'actant **I** potentiel de PLAINTE.

Oper$_1$(*cri*) = *pousser* [ART ~]
Oper$_1$(*grippe*) = *avoir* [ART ~]
Oper$_1$(*ordre*) = *donner* [ART ~ à N]
Oper$_1$(*suprématie*) = *détenir, exercer* [ART ~]

5. Nous ne pouvons malheureusement pas formuler ici le concept crucial d'actant SyntP : cela exigerait une présentation trop longue de la composante SyntP du modèle linguistique Sens-Texte, ainsi que du modèle lui-même. Une fois de plus, il me faut renvoyer le lecteur aux **Références**.

Oper$_1$(*désespoir*) = *être* [*au* ~]
Oper$_1$(*efforts*) = *déployer* [ART ~*s*]
Oper$_1$(*regard*) = *jeter* [ART ~ *à* N] [*lui jeter un regard...*]
Oper$_1$(*précaution*) = *prendre* [ART ~]

N.B. : L'expression entre crochets qui suit la valeur de **Oper**$_1$ (ainsi que de n'importe quelle FL présentée plus loin) est le ***Régime*** de l'élément en question. Le tilde « ~ » remplace le mot clé, et le symbole ART signifie qu'un déterminant (un article, un adjectif possessif ou démonstratif, etc.) doit être utilisé selon les règles de la grammaire française.

Oper$_1$ représente une famille de verbes qui ont reçu le nom de VERBES SUPPORTS (ou « verbes opérateurs ») dans les travaux de M. Gross et de son équipe (Giry-Schneider 1978, M. Gross 1981 et G. Gross 1990, où l'on trouve d'autres références) ; voir ci-dessous, , n^{os} 31-33

Une propriété fort importante des FL standard simples réside dans leur CARACTÈRE UNIVERSEL : elles sont valables pour toutes les langues et sont suffisantes pour la description de la dérivation, de la cooccurrence lexicale restreinte et de la paraphrase dans la grande majorité des cas. Après avoir présenté la liste des FL dans la section qui suit, nous donnerons quelques exemples de FL dans diverses langues naturelles.

Les fonctions lexicales standard simples

Les FL jouent un double rôle dans la description linguistique :

● D'une part, les FL servent à décrire les RELATIONS LEXICALES dans le vocabulaire d'une langue : les relations paradigmatiques et les relations syntagmatiques entre les lexies. Les premières décrivent la dérivation et les phénomènes apparentés, tandis que les secondes spécifient la cooccurrence lexicale restreinte ; voir plus loin.

● D'autre part, les FL servent à décrire la SYNONYMIE ENTRE PHRASES basée sur les relations sémantiques entre lexies, c'est-à-dire les PARAPHRASES LEXICALES (voir Mel'čuk 1988c, 1992).

Le premier aspect concerne directement le dictionnaire de langue : il représente un problème lexicographique avant tout. Le deuxième aspect relève de la sémantique et de la syntaxe profonde de la langue ; il ne fera pas ici l'objet d'un développement particulier.

Sans entrer dans les détails concernant les FL (voir Mel'čuk 1982 et Mel'čuk *et al.* 1984 : 6-13, 1992 : 127-131), nous nous limiterons à la présentation d'une liste des FL (leur ordre dans cette liste correspond à leur ordre d'apparition dans un article de dictionnaire du DEC), accompagnée de brefs commentaires.

N.B. : Les numéros identifiant les différents lexèmes dans les exemples sont empruntés aux entrées publiées ou en préparation du DEC du français contemporain (Mel'čuk *et al.* 1984, 1988, 1992).

Liste de fonctions lexicales standard simples

Pour faciliter la lecture de la présente liste, les FL sont divisées, ainsi qu'à l'accoutumée, en deux classes majeures :

les FL PARADIGMATIQUES, qui représentent les relations paradigmatiques entre lexies et qui couvrent tous les corrélats « dérivationnels » (au sens large et vague) d'une lexie donnée L ;

les FL SYNTAGMATIQUES, qui représentent les relations syntagmatiques entre lexies et qui couvrent tous les corrélats collocationnels de la lexie L.

En règle générale, un élément de la valeur d'une FL paradigmatique est utilisé dans le texte AU LIEU DE son mot clé ; un élément de la valeur d'une FL syntagmatique est habituellement utilisé À CÔTÉ DE (= AVEC) son mot clé[6]. Cependant, cette distinction syntaxique entre les deux classes de FL ne fait que refléter, bien que pas toujours de façon cohérente, une distinction sémantique plus profonde :

- Les FL paradigmatiques visent la NOMINATION/SÉLECTION ; elles doivent nous aider à répondre à des questions du type « Comment appelle-t-on l'objet <la situation> **X**, apparenté<e> à Y ? » – quand on veut parler de **X**, ET NON PAS de Y.

- Les FL syntagmatiques visent la COMBINATOIRE ; elles doivent nous aider à répondre à des questions du type « Comment appelle-t-on l'action <la caractéristique, l'attribut > **X** de Y ? » – quand on veut parler de **Y** ET de X en même temps.

À l'intérieur des deux divisions majeures qui viennent d'être indiquées, nous allons regrouper les FL (là où c'est possible) par la partie du discours de leur valeur : FL nominales, FL adjectivales, FL verbales et FL adverbiales.

Fonctions lexicales paradigmatiques

1. Syn, $\mathbf{Syn}_{\supset}$, $\mathbf{Syn}_{\subset}$, $\mathbf{Syn}_{\cap}$ [synonyme exact et quasi-synonymes ; l'indice $\supset$ signifie ‹sens plus spécifique = plus riche›, l'indice $\subset$, ‹sens moins spécifique = plus général›, tandis que l'indice $\cap$ dénote une intersection des sens] :

Syn(*avion*) = *appareil* $\quad$ $\mathbf{Syn}_{\supset}$ (*respect**I***) = *vénération*
$\mathbf{Syn}_{\subset}$ (*vénération*) = *respect**I*** $\quad$ $\mathbf{Syn}_{\cap}$ (*seconder*) = *assister**II***

N.B. : Les indices ensemblistes s'utilisent avec d'autres FL, toujours dans le même sens[7].

2. $\mathbf{Conv}_{ijk}$[conversif, c'est-à-dire un lexème qui dénote une relation converse de la re-

6. Cette régularité peut être violée, ce qui est systématiquement indiqué dans les entrées lexicales correspondantes : nous faisons ici référence à ce que nous appelons des ***valeurs fusionnées*** de FL, qui sont marquées par deux barres inclinées //. Cependant, nous ne pouvons pas nous étendre davantage sur ce point, si important soit-il.

7. Dans les publications précédentes, les indices ensemblistes d'inclusion $\supset$ et$\subset$ étaient utilisés dans un sens opposé : $\supset$ au sens de ‹plus large›, et$\subset$, au sens de ‹plus étroit›. Nous préférons changer cette pratique parce que ces anciennes interprétations étaient extensionnelles, plutôt que sémantiques pures ; ils visaient l'inclusion des ensembles de référents, et non pas celles de définitions.

lation exprimée par le mot clé de cette FL ; les indices montrent l'ordre des actants SyntP associés au conversif relativement à l'ordre des actants SyntP associés au mot clé du conversif, qui est toujours « 123 »] :

Conv$_{21}$(*craindre*) = *effrayer* [*J'en crains les conséquences* = *Les conséquences de cela m'effraient*]
Conv$_{21}$(*plus*) = *moins* **Conv**$_{3214\cap}$(*vendre*) = *acheter*

3. Anti, Anti$_{\supset}$, **Anti**$_{\subset}$, **Anti**$_{\cap}$ [antonyme exact et quasi-antonymes] :

Anti(*respect*) = *irrespect* **Anti**$_{\subset}$ (*mépris**I***) = *respect**I***
Anti$_{\supset}$(*désespoir*) = *espoir* **Anti**$_{\cap}$ (*aider*) = *gêner*

Anti se combine facilement avec d'autres FL (surtout, avec **Magn, Bon, Ver, Real**) pour former des FL complexes : **AntiMagn**(*majorité*) = *courte, faible* ; **AntiBon**(*choix*) = *malheureux* ; **AntiVer**(*reprocher*) = *à tort* ; **AntiReal**$_3$(*ordre*) = *défier.*

4. Contr [terme contrastif] :

Contr(*d'acier*) = *de velours* [*un regard d'acier* vs *des yeux de velours*]
Contr(*mer **I.1***) = *terre* **Contr**(*tête**I.4***) = *cœur**I.4a***

5. Epit [épithète courante sémantiquement vide] :

Epit(*océan*) = *immense* **Epit**(*gagnant*) = *heureux* **Epit**(*défier*) = *ouvertement*

6. Gener [mot générique pour L qui peut apparaître au moins dans une des deux constructions suivantes] :

1) ⁽**Gener**(L) a(ATTR,--→, DER(L)⁾ = ⁽L⁾ [où DER est un dérivé syntaxique, voir le groupe des FL au n° 8 ci-dessous] ;

2) énumérations du type X_1, X_2,... *et autres* **Gener**(X)] :

Gener(*colère**I***) = *sentiment* [*de* ~] **Gener**(*république*) = *état* [*républicain*]
Gener(*pistolet*) = *arme à feu* [*fusils, pistolets et autres armes à feu*]

7. Figur [métaphore codifiée par la langue dont la combinaison avec le mot clé est un synonyme (plus spécifique) du mot clé] :

Figur(*fumée*) = *rideau* [*de* ~] **Figur**(*haine**I***) = *feu* [*de la* ~]
Figur(*jalousie*) = *démon* [*de la* ~]

8. S$_0$, V$_0$, A$_0$, Adv$_0$ [dérivés syntaxiques : nom, verbe, adjectif, adverbe, dérivés du mot clé sans changement du sens, de sorte que ⁽**S$_0$**(X)⁾ = ⁽X⁾] :

S$_0$(*acheter**I***) = *achat**Ia*** **A$_0$**(*école**I.1a***) = *scolaire*
V$_0$(*promesse**I***) = *promettre**I*** **Adv$_0$**(*honnête*) = *honnêtement*

Fonctions nominales

9. $\mathbf{S_1}$, $\mathbf{S_2}$, $\mathbf{S_3}$, ...[nom typique pour le premier, deuxième, troisième, ... actant du mot clé] :

$\mathbf{S_1}$(*crime*) = *criminel* $\mathbf{S_2}$(*crime*) = *victime* [*de* ART ~]
$\mathbf{S_1}$(*acheter1*) = *acheteur* $\mathbf{S_{2\supset}}$(*acheter1*) = *marchandise*
$\mathbf{S_3}$(*acheter1*) = *vendeur* $\mathbf{S_4}$(*acheter1*) = *prix*

10. $\mathbf{S_{instr}}$, $\mathbf{S_{loc}}$, $\mathbf{S_{med}}$, $\mathbf{S_{mod}}$, $\mathbf{S_{res}}$ [nom typique pour le circonstant exprimant l'instrument, le lieu, le moyen, le mode et le résultat] :

$\mathbf{S_{instr}}$(*peindre*) = *pinceau, brosse* $\mathbf{S_{loc}}$(*hostilités*) = *théâtre* [*des* ~]
$\mathbf{S_{med}}$(*peindre*) = *peinture*
$\mathbf{S_{mod}}$(*écrire*) = *écriture* $\mathbf{S_{res}}$ (*copier*) = *copie*

11. Sing [‘un quantum régulier de ...’] :

Sing(*flotte*) = *navire* **Sing**(*ail*) = *gousse* [*d'*~] ; *tête* [*d'*~] **Sing**(*riz*) = *grain* [*de* ~]

12. Mult [‘ensemble régulier de ...’] :

Mult(*navire*) = *flotte* **Mult**(*chien*) = *meute* **Mult**(*barbares*) = *horde*
Mult(*abeille*) = *essaim, nuée* **Mult**(*oiseau*) = *volée*

13. Cap [‘chef de ...’] :

Cap(*université*) = *recteur* **Cap**(*avion*) = *commandant*
Cap(*théâtre*) = *directeur*

14. Equip [‘équipe de ...’] :

Equip(*navire, avion*) = *équipage* **Equip**(*théâtre*) = *troupe*

15. Centr [‘le centre de ...’, ‘le milieu de ...’] :

Centr(*problème*) = *cœur* [*du* ~]

Cette FL est souvent utilisée dans des FL complexes, comme, par exemple :

$\mathbf{Loc_{in}Centr}$(*hiver, nuit, mêlée*) = *au cœur* [*de* ART ~]

Pour $\mathbf{Loc_{in}}$, voir plus loin, n° 28.

16. Culm [‘culmination de ...’] :

Culm(*joie1*) = *comble**II*** [*de la* ~] **Culm**(*colère1*) = *paroxysme* [*de la* ~]

Fonctions adjectivales

17. $\mathbf{A_1}$, $\mathbf{A_2}$, ... [modificateur typique pour le premier, deuxième, ... actant du mot clé] :

$\mathbf{A}_1$(*mépris*) = *plein, rempli* [*de* ~] $\mathbf{A}_2$(*mépris*) = *couvert* [*de* ~]
$\mathbf{A}_1$(*chercher*) = *//en quête* [*de* N] $\mathbf{A}_2$(*diriger*) = *//sous la direction* [*de* N]

18. Able$_1$, Able$_2$, ... [‘tel qu'il peut ...’, ‘tel qu'on peut le ...’, etc.] :

Able$_1$(*peur*) = *peureux* **Able$_2$**(*peur*) = *effrayant*

19. Qual$_i$ [‘tel qu'il entraîne Able$_i$ avec une forte probabilité’] :

Qual$_1$(*tromper*) = *malhonnête* **Qual$_2$**(*tromper*) = *naïf*

Fonctions lexicales syntagmatiques

Fonctions adjectivales

20. Magn [‘très’, ‘intense/intensément’, ‘à un degré élevé’] :

Magn(*mémoire*) = *prodigieuse, excellente, étonnante, d'éléphant*
Magn(*bruit*) = *infernal, de tonnerre*
Magn(*remercier*) = *vivement, chaleureusement, de tout cœur ; infininiment* | seulement performatif

21. Plus, Minus [‘plus’, ‘moins’ ; ces FL ne s'emploient qu'en combinaison avec d'autres FL] :

IncepPredPlus(*joie**I***) = *grandir* **IncepPredMinus**(*joie**I***) = *faiblir*
[pour **Incep** et **Pred**, voir n^{os} 35 et 30]

IncepPredPlus(*vent**I**.1*) = *augmenter, prendre de la force, s'élever*
IncepPredMinus(*vent**I**.1*) = *se calmer, mollir*

Ainsi que nous l'avons dit plus haut, nous n'expliquerons pas la structure syntaxique des FL complexes.

22. Ver [‘tel qu'il doit être’, ‘correct’] :

Ver(*peur*) = *justifiée* **Ver**(*appareil*) = *exact, précis* **Ver**(*proposition*) = *sérieuse*

23. Bon [‘bon’ – expression qu'on emploie comme une louange standard codifiée par la langue] :

Bon$_2$(*conseil**I**.1*) = *précieux* **Bon**(*compliment*) = *recherché, bien tourné*
Bon(*temps*) = *beau*

24. Pejor [‘pire’ = **MinusBon**] :

CausPredPejor(*joie**I***) = *altérer, gâcher* [*la joie**I***]

N.B. : IncepPredPejor = Degrad, voir n° 46 ; **CausPredPejor**(*joie1*) peut donc être réécrit comme **CausDegrad**(*joie1*).

25. Pos_1, Pos_2, ...[‘évaluation positive’ – expression qu'on emploie comme expression standard de l'évaluation positive du premier, deuxième, ... actant SyntP du mot clé] :

Pos_2(*opinion*) = *favorable, avantageuse* **Pos_2**(*critique*) = *favorable, élogieuse*

Fonctions adverbiales

26. Adv_1, Adv_2, ...[adverbe typique pour caractériser le comportement du premier, deuxième, ... actant du mot clé ; autrement dit, l'adverbe qui signifie ‘étant le premier, deuxième, ... actant de’] :

Adv_1(*joie1*) = *avec* [~] **Adv_2**(*joie1*) = *à* [*la grande ~ de* N]

27. Instr [préposition régissant le mot clé et signifiant ‘au moyen de’] :

Instr(*mainsI.a*) = *de, avec* [*les ~*] ; *à* [*la ~*] | avec les expressions signifiant ‘fabrication’)
Instr(*tête*) = *de, avec* [*la ~*]
Instr(*machine à écrire*) = *à* [*la ~*]

28. Loc_{in}, Loc_{ab}, Loc_{ad} [préposition régissant le mot clé et signifiant ‘se trouvant [de façon spatiale ou temporelle] dans ...’ (**Loc_{in}**), ‘se déplaçant à partir de’ (**Loc_{ab}**), ‘se déplaçant vers l'intérieur de’ (**Loc_{ad}**)] :

$Loc_{in/ad}$(*gare*) = *à* [ART ~]
Loc_{in}(*personnel*) = *au sein de* [*le ~*] [*au sein du personnel*]
Loc_{ab}(*gare*) = *de* [ART ~] ; **Loc_{ab}**(*1970*) = *depuis* [~].

29. Propt [préposition régissant le mot clé et signifiant ‘à cause de’] :

Propt(*peur*) = *de, par* [~] **Propt**(*respect*) = *par* [~]
Propt(*maladie*) = *pour cause de* [~]

Fonctions verbales

30. Pred [‘être’ ; verbalisateur des FL adjectivales] ; n'apparaît qu'en combinaison avec d'autres FL : voir les exemples donnés pour les FL **Plus** et **Minus**, n° 21.

Passons maintenant aux fonctions verbales mettant en jeu les trois éléments syntaxiques de surface [= SyntS] qui jouent un rôle central dans la phrase : le S(ujet) G(rammatical), le C(omplément d')O(bjet)$^{dir(ect)}$ et le CO^{indir}. Ces FL se présentent par TRIPLETS, fait qu'explique leur rôle syntaxique : elles servent à relier TROIS noms, à savoir la lexie vedette C_0 (en tant que mot clé) avec ses A(ctants) SyntP **I** et **II**.

31-33. Le premier triplet est formé par les FL **$Oper_i$**, **$Func_i$** et **$Labor_{ij}$**, qui formalisent la notion assez connue de *verbe support*. Ces FL (et à l'évidence les éléments de leurs

valeurs) sont des verbes sémantiquement vides (ou vidés de leur sens dans le contexte de leur mot clé), qui servent à « verbaliser » les noms prédicatifs, en exprimant le mode et le temps. Leur vocation est, avant tout, syntaxique[8] et ils sont distingués uniquement par le rôle syntaxique du mot clé et des actants syntaxiques de celui-ci :

● La FL elle-même est déterminée par le rôle syntaxique du mot clé : **Oper$_i$** prend son mot clé en tant que CO^{dir} (*faire* UNE ERREUR, *recevoir* UN ORDRE), **Func$_i$** prend son mot clé en tant que SG (CET ORDRE *vient de ...*, CET ORDRE *vous concerne*), et **Labor$_{ij}$** prend son mot clé en tant que CO^{indir} (*soumettre ... à* UNE ANALYSE, *prendre ... en* LOCATION).

● L'indice actanciel d'une FL du type considéré est déterminé par le rôle des actants du mot clé : l'indice $_1$ renvoie à l'actant SyntP **I** du mot clé, $_2$ à l'actant SyntP **II**, $_3$ à l'actant SyntP **III**, etc.

Rôle syntaxique de surface du mot clé Fonction lexicale	**SG**	**CO^{dir}**	**CO^{indir}**
Oper$_{1/2}$	**I / II**	C_0	**II / I**
Func$_{0/1/2}$	C_0	— **/I / II**	**II/ I**
Labor$_{12/21}$	**I / II**	**II / I**	C_0

Oper$_1$(*attention*) = *prêter* [~ *à* N] **Oper$_2$**(*attention*) = *attirer* [ART ~]
Oper$_1$(*conseil*) = *donner* [ART/ø ~ *à* N]
Oper$_2$(*conseil*) = *recevoir* [ART/ø ~ *de* N]
Func$_0$(*silence*) = *règne* [Loc$_{in}$ N]

[L'indice zéro indique que ce verbe n'a pas de complément : c'est un intransitif « absolu ». Notons encore que, dans nos exemples, les verbes sont à l'infinitif si le mot clé est un complément et à l'indicatif présent de la 3^e personne, si le mot clé est le sujet grammatical.]
Func$_1$(*aide*) = *vient, provient* [*de* N] **Func$_2$**(*liste*) = *contient, comprend* [N]
Func$_2$(*danger*) = *menace* [N]
Labor$_{12}$(*traitement*) = *soumettre* [N *à* ART ~]
Labor$_{12}$(*soin*) = *entourer* [N *de* (A$_{poss}$)~*s*]

34-36. Le deuxième triplet comprend les FL **Incep, Fin** et **Cont**, qui expriment les trois PHASES différentes d'un état ou d'un événement : le début, la fin et la continuation. Ces FL, qu'on peut appeler ***phasiques***, sont donc des verbes sémantiquement pleins qui ont les significations suivantes :

8. Cela ne signifie pas que ces verbes ne jouent aucun rôle sémantique : ils sont cruciaux pour exprimer des perspectives COMMUNICATIVES différentes. Les phrases *Le général St-Germain a donné un ordre au lieutenant Polguère* et *Le lieutenant Polguère a reçu du général St-Germain un ordre* décrivent la même situation et ont la même structure sémantique, mais constituent deux messages différents. Ce sujet a été mentionné à la page 198.

Incep(P) = ‘commencer à P-er[faire l’action P]’,
Fin(P) = **Incep**(**non**P) = ‘cesser de P-er’ = ‘commencer à ne pas P-er’
Cont(P) = **non Fin**(P) = **non Incep**(**non**P) = ‘continuer de P-er’ = ‘ne pas cesser de P-er) = ‘ne pas commencer à ne pas P-er’

Étant donné leur sens, les FL **Incep**, **Fin** et **Cont** doivent prendre des verbes pour mot clé. Cependant, l’application de ces FL aux verbes français ne présente aucun intérêt du point de vue lexicographique puisque, avec un verbe, ces fonctions sont presque toujours exprimées de la façon régulière indiquée ci-dessus : **Incep**(*chanter*) = *commencer* [*à* ~], **Fin**(*lire*) = *cesser* [*de* ~], etc. Il existe quand même quelques cas particuliers : **Incep**(*dormir*) = *s’endormir* ; **Fin**(*dormir*) = *se réveiller* ; **Incep**(*exister*) = *naître*, **Fin**(*exister*) = *mourir*, etc. Mais comme ils sont peu fréquents, nous n’en tiendrons pas compte.

Par contre, l’application de ces trois FL aux noms prédicatifs retourne des valeurs aussi riches que variées en français. Notons, cependant, qu’elles doivent s’exprimer inévitablement en combinaison avec d’autres FL. En fait, du point de vue sémantique, **Incep**, **Fin** et **Cont** sont des prédicats MONOactanciels (un événement/acte/processus ... commence/cesse/continue) ; pour cette raison, ces trois FL n’ont pas de structure actancielle propre et ont donc besoin de s’appuyer sur les FL actancielles, telles que $\mathbf{Oper}_i$, $\mathbf{Func}_i$ et $\mathbf{Labor}_{ij}$, d’une part, et $\mathbf{Real}_i$, $\mathbf{Fact}_i$ et $\mathbf{Labreal}_{ij}$, de l’autre.

$\mathbf{IncepOper}_1$(*suprématie*) = *arriver* [*à* ART ~], *acquérir, obtenir* [ART ~]
$\mathbf{FinOper}_1$(*suprématie*) = *perdre* [ART ~] $\mathbf{FinOper}_1$(*influence*) = *perdre* [ART ~]
$\mathbf{ContOper}_1$(*suprématie*) = *conserver, garder* [ART ~]
$\mathbf{ContOper}_1$(*influence*) = *garder* [ART ~]
$\mathbf{IncepOper}_1$(*caractère*) = *revêtir* [ART ~] $\mathbf{IncepOper}_1$(*forme*) = *prendre* [ART ~]
$\mathbf{IncepOper}_1$(*feu* [tir]) = *ouvrir* [ART ~ *sur* N]
$\mathbf{IncepOper}_2$(*feu* [tir]) = *se trouver* [*sous* ART ~]
$\mathbf{ContOper}_2$(*feu* [tir]) = *rester* [*sous* ART ~]
$\mathbf{IncepFunc}_0$(*vent*) = *se lève* $\mathbf{FinFunc}_0$(*vent*) = *se calme*
$\mathbf{IncepFunc}_1$(*colère*) = *s’empare* [*de* N]
$\mathbf{IncepFact}_0$(*film*) = *sort* (*sur les écrans*)

37-39. Le troisième triplet comprend les FL **Caus, Liqu** et **Perm**, qui expriment les trois types de causation d’un état ou d’un événement. Ces FL, qu’on pourrait qualifier de ***causatives***, sont donc des verbes sémantiquement pleins qui ont les significations suivantes :

Caus(P) = ‘causer que P [faire en sorte que P a lieu]’
Liqu(P) = **Caus**(**non**P) = ‘liquider P’ = ‘causer que non P’
Perm(P) = **nonLiqu**(P) = **nonCaus**(**non**P) = ‘permettre P’ = ‘ne pas liquider P’ = ‘ne pas causer que non P’

Pour bien comprendre l’usage des FL causatives, il nous faut toucher à deux aspects de leur comportement : d’une part, la structure actancielle des FL complexes dans lesquelles elles sont mises en jeu, et d’autre part, le lien entre les FL causatives et les sens phasiques, c’est-à-dire les FL **Incep, Fin** et **Cont**.

Les FL causatives et la structure actancielle des FL complexes. À la différence

des autres FL qui ne changent jamais la structure actancielle de la situation décrite par la lexie vedette, une FL causative introduit, en règle générale, un nouvel actant : le causateur. Celui-ci est exprimé comme l'actant SyntP [= ASyntP] **I** de la FL causative, et par conséquent, les actants de départ de la lexie vedette sont tous décalés. On le voit très bien dans l'exemple banal de la construction causative française, où l'ASyntP **I** de départ devient l'ASyntP **III** :

(6) *Jean* [= **I**] *écrit une lettre* [= **II**] vs *Colette* [= **I**] *fait écrire une lettre* [= **II**] *à Jean* [= **III**].

Le décalage des ASyntP provoqué par une FL causative s'exprime au moyen des FL verbales vides, soit $\mathbf{Oper}_i$, $\mathbf{Func}_i$ et $\mathbf{Labor}_{ij}$, c'est-à-dire que nous faisons intervenir les combinaisons du type $\mathbf{CausOper}_i$, $\mathbf{LiquFunc}_i$, etc. Pour bien présenter la systématicité de cette description, je vais analyser un exemple de façon détaillée. Soit une lexie vedette : ENVIE, au sens de 'désir causé par un besoin'. Elle a deux actants SyntP : **I** – celui qui ressent l'envie, et **II** – l'objet de l'envie, comme dans la phrase (7) :

(7) *Pierre* [= **I**] *a* [= $\mathbf{Oper}_1$] ENVIE *d'y aller* [= **II**].

Cette phrase peut être « enchâssée » dans une expression causative, par exemple, (8) :

(8) **a**. *Cette aventure* ***a privé*** *Pierre de l'*ENVIE *d'y aller.*
b. *Cette aventure* ***a ôté*** *à Pierre l'*ENVIE *d'y aller.*

Ces deux phrases sont équivalentes quant à leur contenu propositionnel, et les verbes PRIVER et ÔTER sont clairement des FL de ENVIE. Mais comment les décrire par le symbolisme des FL ?

Sémantiquement, ces deux verbes veulent dire la même chose : 'CAUSER LA NON-EXISTENCE [de l'envie de Pierre d'y aller]' ; ce sens doit être exprimé par la FL **Liqu**.

Syntaxiquement, cependant, les deux verbes diffèrent par leur régime, et pour exprimer cette différence, nous avons besoin des FL du type de $\mathbf{Oper}_i$, $\mathbf{Func}_i$ et $\mathbf{Labor}_{ij}$. Plus précisément, *priver Pierre de l'envie d'y aller* s'interprète comme 'causer que Pierre n'**a** plus d'envie' ; comme *a* [= AVOIR] est $\mathbf{Oper}_1$ de ENVIE, *priver* s'écrit $\mathbf{LiquOper}_1$(*envie*). A son tour, *ôter à Pierre l'envie d'y aller* s'interprète comme 'causer que l'envie n'**est** plus à Pierre' ; *est à* [= ÊTRE *à*] est $\mathbf{Func}_1$ de ENVIE (non admissible comme tel en français), de sorte que *ôter* s'écrit comme $\mathbf{LiquFunc}_1$(*envie*). De façon similaire, dans *plonger Suzanne dans une rage froide*, le verbe *plonger* est $\mathbf{CausOper}_1$(*rage*), tandis que le verbe *ouvrir* dans *lui ouvrir une perspective* est $\mathbf{CausFunc}_1$(*perspective*). Ajoutons encore les exemples suivants :

$\mathbf{CausOper}_1$(*désespoir*) = *pousser, réduire* [N *au* ~], *jeter* [N *dans le* ~], *frapper* [N *de* ~]
$\mathbf{CausFunc}_0$(*difficulté*) = *créer, poser* [*une* ~] $\mathbf{LiquFunc}_0$(*assemblée*) = *dissoudre* [ART ~]
$\mathbf{ContOper}_1$(*suprématie*) = *maintenir* [ART ~][9]

9. On notera la possibilité d'un emploi différent du verbe MAINTENIR avec SUPRÉMATIE, fréquent avec d'autres noms prédicatifs :
(i) *Son prestige lui maintient sa suprématie sur ses collègues*, où MAINTENIR réalise une autre FL de SUPRÉMATIE : $\mathbf{CausFunc}_1$.

LiquFunc$_2$(*attention*) = *détourner* [l'~ *de* N *de* N]
Caus$_2$Func$_2$(*attention*) = *accaparer* [*l'*~]
Perm$_1$Fact$_0$(*colère*) = *s'abandonner* [*à la* ~]

Les FL causatives elles-mêmes peuvent ne pas avoir d'indice actanciel : c'est le cas général, où la FL du type **Caus** introduit un actant supplémentaire vis-à-vis des actants de la lexie vedette. Cependant, il est également possible que le causateur soit un des actants de la lexie vedette ; alors il est indiqué par l'indice actanciel correspondant ; voir **Caus$_2$Func$_2$**(*attention*) et **Perm$_1$Fact$_0$**(*colère*) ci-dessus, où le causateur est en même temps un des actants du mot clé.

Le lien entre les FL causatives et les FL phasiques. Comme la causation est intimement liée à la PHASE du fait causé (on cause soit le commencement, soit la continuation, soit la cessation d'un procès, d'un événement, etc.), nous devrions, pour être rigoureux, toujours indiquer, après une FL causative, la FL phasique correspondante. Cependant, pour alléger l'écriture, nous adopterons la convention suivante :

> Étant donné que le cas le plus courant est la causation du commencement du fait causé, au lieu de **CausIncep**, nous écrirons simplement **Caus**.

Par contre, les autres combinaisons « causation-phase » devront toujours être explicitement marquées. Ainsi écrirons-nous **CausCont** pour indiquer la continuation, et **CausFin** pour marquer la cessation ; cependant, **CausFin** est obligatoirement remplacé par la FL **Liqu**, qui est par définition son équivalent. Par conséquent, bien que pour *mettre* N *sous la forme de* ... nous ayons dû écrire **CausIncepOper$_1$**(*forme*), nous écrirons, en simplifiant, **CausOper$_1$**(*forme*) ; de même, ÉTABLIR, qui dans *établir la paix* est strictement parlant **CausIncepFunc$_0$**(*paix*), apparaîtra, d'après la convention retenue, comme **CausFunc$_0$**(*paix*). En même temps, par exemple, *maintenir la paix* devra obligatoirement s'écrire au complet : **CausContFunc$_0$**(*paix*).

40-42. Le quatrième triplet comprend les FL **Real$_i$**, **Fact$_i$** et **Labreal$_{ij}$**, qui expriment *grosso modo* le sens ‘réaliser les « objectifs » inhérents de la chose [désignée par le mot clé]’. Ces FL sont donc des verbes sémantiquement pleins. Leur syntaxe est cependant identique à celle des FL **Oper$_i$**, **Func$_i$** et **Labor$_{ij}$**, de sorte que **Real$_i$** correspond à **Oper$_i$**, **Fact$_i$** à **Func$_i$**, et **Labreal$_{ij}$** à **Labor$_{ij}$**. Ainsi, **Real$_i$** prend le mot clé comme son actant **II** [= CO^{dir}], **Fact$_i$**, comme son actant **I** [= SG], et **Labreal$_{ij}$** comme son actant **III** [= CO^{indir}] ; les indices actanciels sont déterminés de la même façon que ci-dessus : **Real$_i$** a en tant que SG l'actant SyntP **I** du mot clé, le SG de **Real$_2$** est l'actant SyntP **II** du mot clé, etc.

Real$_1$(*peine* [jurid.]) = *imposer, infliger* [ART ~ *à* N]
Real$_2$(*peine* [jurid.]) = *purger* [ART ~]
Real$_2$(*piège*) = *donner, tomber* [*dans* ART ~]
Real$_3$(*ordre*) = *exécuter* [ART ~]
Real$_1$(*film*) = *jouer* [ART ~] [*On joue ce film à l'Odéon*]
Fact$_0$(*film*) = *est à l'affiche* [*Ce film est à l'affiche à l'Odéon*]
Fact$_0$(*rêve*) = *se réalise* (voir angl. *His dream came true*)

43. Involv [‘affecter’, ‘toucher’ : verbe qui prend le mot clé comme son SG et le nom signifiant l'objet qui subit l'action de la situation désignée par le mot clé, sans en être un participant, comme son CO principal] :

Involv(*vent*) = *agite, secoue* [*un arbre*] · *cingle, brûle* [*le visage de Pierre*] ; *plie, incline, courbe* [*les roseaux*] , ...
Involv(*odeur*) = *remplit* [*la pièce*]
Involv(*lumière*) = *se diffuse1, se répand* [*dans la pièce*]

44. Manif ['se manifester dans ...'] : le mot clé est le SG] :

Manif(*joie1*) = *éclate, jaillit* **Manif**(*colère*) = *éclate, explose*

Très souvent, **Manif** apparaît avec $\mathbf{Caus}_1$, avec laquelle elle forme une FL complexe :

$\mathbf{Caus}_1\mathbf{Manif}$(*excuse*) = *présenter* [*ses* ~]
$\mathbf{Caus}_1\mathbf{Manif}$(*opinion*) = *exprimer, formuler* [ART ~]

Les deux FL suivantes – **Prepar** et **Prox** – n'ont pas de structure actancielle propre et n'apparaissent qu'avec les FL du type $\mathbf{Oper}_i$ ou $\mathbf{Real}_i$, c'est-à-dire qu'elles sont toujours utilisées dans des FL complexes.

45. Prepar ['préparer N pour ...'] :

$\mathbf{PreparFact}_0$(*fusil*) = *charger* [ART ~] $\mathbf{PreparFact}_0$(*voiture*) = *// faire le plein*

46. Prox ['être sur le point /être prêt de ...'] :

$\mathbf{ProxOper}_1$(*désespoir1*) = *être au bord <à la limite>* [*du* ~]
$\mathbf{ProxFunc}_0$(*orage1*) = *(s')approche*

47. Degrad ['se dégrader', 'devenir pire' ; le mot clé est le SG] :

Degrad(*cœur1.1a*) = *faiblit* **Degrad**(*lait*) = *tourne* **Degrad**(*vin*) = *s'aigrit*

48. Son ['émettre le son typique' ; le mot clé est le SG] :

Son(*plancher*) = *craque* **Son**(*chien*) = *aboie* **Son**(*moteur*) = *ronronne, vrombit*

49. Imper [formule exclamative qui exprime l'ordre, la prière, etc. autrement que par une forme impérative régulière du verbe] :

Imper(*secourir*) = *Au secours !* **Imper**(*tirer*) = *Feu !*

50. Result [verbe désignant 'l'état qui résulte d'un événement' ; l'indice actanciel spécifie l'actant du mot clé qui est le SG de **Result**] :

$\mathbf{Result}_1$(*avoir promis1*) = *//est lié par* A_{poss} *promesse*
$\mathbf{Result}_2$(*avoir promettre1*) = *//a la promesse de* N

Les trois FL qui suivent – **Obstr, Stop** et **Excess** – prennent, par défaut, le mot clé comme SG. Au cas où leur SG doit être la désignation de la personne (= ASyntP **I** du mot clé), cela est indiqué par l'indice 1.

51. Obstr ['fonctionner avec difficulté'] :

Obstr(*souffle*) = [*lui*] *manque* **Obstr**(*vue*) = *se brouille*

52. Stop ['arrêter de fonctionner'] :

Stop$_1$(*souffle*) = *perdre* [*le* ~]
Stop(*cœur**I.1a***) = *s'arrête, flanche* **Stop**(*cœur**I.4a***) = *se brise, se rompt*

53. Excess ['fonctionner d'une façon anormalement excessive'] :

Excess(*cœur**I.1a***) = *palpite, accélère* **Excess**(*moteur*) = *s'emballe*

54. Sympt [expression verbale complexe signifiant un symptôme physique d'une émotion, d'un état, etc., qui est un état particulier d'une partie du corps ou d'un organe]. **Sympt** prend trois actants, les indices actanciels étant attribués comme suit : l'indice 1 correspond à la personne « propriétaire » de la partie du corps impliquée et sujet de l'émotion ou de l'état en question, l'indice 2, à la partie du corps, et l'indice 3, à l'émotion. Conformément à notre usage d'exprimer la conversion syntaxique, l'ordre des indices actanciels signale leur rôle syntaxique de surface : l'indice venant en premier correspond au SG de l'expression de **Sympt**, celui venant en deuxième, à son CO principal et celui venant en dernier à son CO secondaire.

Cette FL est utilisée nécessairement dans des combinaisons avec **Obstr**, **Stop** et **Excess** (ainsi qu'avec d'autres FL non standard) :

Obstr(*parole*) – **Sympt**$_{13}$(*colère*) = *bafouer, bégayer* [*de colère*]
Stop(*parole*) – **Sympt**$_{13}$(*étonnement*) = *être muet* [*d'étonnement*]
Excess(*dents*) – **Sympt**$_{123}$(*colère*) = *grincer des dents* [*de colère*]
Excess(*tête*) – **Sympt**$_{12}$(*avoir sommeil*) = *dodeliner de la tête*
Excess(*cheveux*) – **Sympt**$_{23}$(*horreur*) = *ses cheveux se dressent* [*d'horreur*]

Pour clore cette section, nous voudrions illustrer le caractère universel des fonctions lexicales par une liste d'exemples de fonctions lexicales dans d'autres langues que le français.

Anglais

Magn(*rain* 'pluie') = **heavy** 'lourd'
Magn(*argument* 'argument') = *strong* 'fort', *weighty* 'pesant'
Magn(*applause* 'applaudissements') = *thunderous* 'de tonnerre', *deafening* 'assourdissants', *boisterous* 'bruyants', *whirl-wind* 'tourbillonnants', ...

Oper$_1$(*trip* 'voyage') = *take* [ART ~] 'prendre'
Oper$_1$(*deal* 'accord, transaction') = *make, strike* [ART ~] 'faire', 'frapper'
Oper$_1$(*apologies* 'excuses') = *offer* [N A_{poss} ~] 'offrir'

Allemand

Magn(*Regen* 'pluie') = *starker* 'fort', *Platz-* 'à éclat'

Magn(*Argument* 'argument') = *gewichtiges* 'pesant', *schlagendes* 'battant', *unschlagbares* 'imbattable'
Magn(*Applaus* 'applaudissements') = *tosender* 'mugissant'

Oper$_1$(*Reise* 'voyage') = [ART ~] *machen* 'faire'
Oper$_1$(*Übereinkunft* 'accord') = [*über* ART ~] *erzielen* 'obtenir'
Oper$_1$(*Entschuldigung* 'excuses') = [N_{dat} A_{poss} ~] *entgegenbringen* 'montrer'

Russe

Magn(*dožd'* 'pluie') = *sil'nyj* 'fort', *prolivnoj* 'd'averse'
Magn(*dovod* 'argument') = *veskij* 'pesant'
Magn(*applodismenty* 'applaudissements') = *burnye* 'tempétueux'

Oper$_1$(*putešestvie* 'voyage') = *soveršit'* [~*e*] 'accomplir'
Oper$_1$(*soglašenie* 'accord') = *zaključit'* [~*e*] 'contracter', *pridti* [*k* ~*ju*] 'arriver à'
Oper$_1$(*izvinenija* 'excuses') = *prinesti* [N_{dat} A_{poss} ~*ja*] 'apporter'

Polonais

Magn(*deszcz* 'pluie') = *silny* 'fort', *mocny* 'puissant', *ulewny* 'd'averse'
Magn(*argument* 'argument') = *silny* 'fort', *mocny* 'puissant'
Magn(*oklaski* 'applaudissements') = *burzliwe* 'tempétueux', *olbrzymie* 'énormes'

IncepOper$_1$(*podróż* 'voyage') = *wybrać się, wyruszyć się* [*w* ~] 'partir en'

[le polonais n'a pas de **Oper**$_1$ pour ce lexème]

Oper$_1$(*porozumienie* 'accord') = *dojść* [*do* ~*a*] 'arriver à'
Oper$_1$(*przeprosiny* 'excuses') = ? [n'existe pas ; on utilise le verbe *przepraszać* 'présenter ses excuses']

Hongrois

Magn(*eső* 'pluie') = *zuhogó* 'torrentiel'
Magn(*érv* 'argument') = *komoly* 'sérieux'
Magn(*taps* 'applaudissements') = *viharos* 'tourbillonnants', *vas-* 'de fer'

Oper$_1$(*utazás* 'voyage') = [~*t*] *tenni* 'faire'
Oper$_1$(*lépés* 'pas') = [~*t*] *tenni* 'faire'
Oper$_1$(*hatalom* 'pouvoir') = [~*t*] *birtokolni* 'posséder'

Arabe

Magn(*maṭar* 'pluie') = *gazīr* 'abondante'
Magn(*ḥuǯǯat* 'argument') = *dāmiga* 'frappante', *qawijja* 'forte'
Magn(*taṣfiq* 'applaudissements') = *ḥārr* 'chaud', *qawijj* 'fort'

Oper$_1$(*safar* 'voyage') = *qāma* [*bi* ~] 'faire'
Oper$_1$(*ʔittifāq* 'accord') = *tawaṣṣala* [*ʔila* ~] 'arriver à', *abrama* [~] 'conclure'

Oper$_1$(*2iʕtiðārāt* ‘excuses’) = *qaddama* [ART~] ‘avancer [trans.]’

Chinois

Magn(*yǔ* ‘pluie’) = *dà* ‘gros’
Magn(*lùnjù* ‘argument’) = *yǒulì-de* ‘ayant de la force’
Magn(*zhǎngshēng* ‘applaudissements’) = *léidòng* ‘de tonnerre’ | postpos

Oper$_1$(*lǚtú* ‘voyage’) = *tàshang* [~] ‘marcher sur’
Oper$_1$(*xiéyí* ‘accord’) = *dáchéng* [~] ‘arriver à’
Oper$_1$(*qiàn* ‘excuse’) = *dào* [*jīge* ‘une’ ~] ‘dire’

Somalien

Oper$_1$(*birmad* ‘attaque’) = *ki ʕi* ‘lever’, *dhufan* [~] ‘frapper’
Oper$_1$(*rajoda* ‘espoir’) = *qabi* [~] ‘tenir’
Oper$_1$(*dagaal* ‘lutte’) = *jiri* [~] ‘se trouver dans’
Oper$_1$(*jawabta* ‘réponse’) = *ʕelin* [~] ‘tourner’
Oper$_1$(*fiiro* ‘attention’) = *lahaan* [~] ‘avoir’, *yeelan* [~] ‘faire’
Oper$_1$(*moqif* ‘position’) = *taagan* [~] ‘être debout dans’

Albanais

Oper$_1$(*besim* ‘confiance’) = *ka* [~] ‘avoir’
Oper$_1$(*be* ‘serment’) = *bën* [~] ‘faire’
IncepOper$_1$(*bela* ‘problèmes’) = *bie* [*në* ~] ‘tomber dans’
Oper$_2$(*qotek* ‘rossée’) = *ha* [~] ‘manger’
Real$_1$(*borxhin* ‘dette’) = *bën* [~] ‘faire’ |= ‘payer une dette’ et non *‘faire une dette’]

Persan

Oper$_1$(*kotak* (rossée)) = [~] *zadan* (frapper)
Oper$_1$(*ġalabe* (victoire)) = [~] *kardan* (faire)
Oper$_1$(*qose* (ennui, tracas) = [~] *xordan* (manger)
Oper$_1$(*fahm* (compréhension)) = [~] *dāštan* (avoir)
Oper$_1$(*kalame* ‘parole(s)’) = [~] *harf zadan* ‘parler’
Real$_1$(*ǰaru* ‘balai’) = [~] *kardan* ‘faire’ [= ‘balayer’]
Real$_1$(*češm* ‘œil’) = [~] *duxtan* ‘coudre’ [= ‘observer’]

En persan, la plupart des sens verbaux (à peu près 90 %) sont exprimés non pas par des lexèmes individuels, mais par des expressions bilexémiques du type illustré ci-dessus, comprenant des fonctions lexicales comme **Oper** et **Real**. On voit que dans cette langue, les FL occupent une place vraiment spéciale !

Remerciements

Le texte de la présente communication a été lu par F. Bélanger, L. Iordanskaja et A. Kharrat ; nous saisissons cette occasion pour leur exprimer toute notre reconnaissance pour leurs remarques et leurs commentaires.

Abréviations et notations

A : actant
ART : déterminant quelconque
C : colonne (du tableau de régime) ; actant syntaxique de surface
DEC : Dictionnaire explicatif et combinatoire
FL : fonction lexicale
G : gouverneur syntaxique
L : lexie
SG : sujet grammatical
SSém : structure sémantique
SyntP : syntaxique profond
SyntS : syntaxique de surface
TST : théorie Sens-Texte
Λ : ensemble vide
~ : lexie vedette
⌜*X+Y+...+Z⌝: phrasème, ou expression figée, qui a une entrée séparée dans le DEC

L : langue donnée.

Références

ABEILLÉ, Anne (1988) : « Light Verb Constructions and Extraction out of NP in Tree Adjoining Grammar », *CLS 24-I*, pp. 1-16.

ALONSO RAMOS, Margarita, TUTIN, Agnès et Guy LAPALME (1992) : « Lexical Functions of *Explanatory Combinatorial Dictionary* for Lexicalization in Text Generation », P. Saint-Dizier et E. Viegas (dir), *Proceedings of the 2nd Seminar on Computational Lexical Semantics*, Toulouse, IRIT, pp. 157-168.

ALONSO RAMOS, Margarita et Agnès TUTIN (1993) : « Les Fonctions Lexicales du *Dictionnaire Explicatif et Combinatoire* pour l'étude de la cohésion lexicale », *Lingvisticae Investigationes*, 17-1.

FONTENELLE, Thierry (1993) : « Using a Bilingual Computerized Dictionary to Retrieve Support Verbs and Combinatorial Information », *Acta Linguistica Hungarica*, 42-1/4.

GIRY-SCHNEIDER, Jacqueline (1978) : *Les nominalisations en français. L'opérateur « faire » dans le lexique*, Genève – Paris, Librairie Droz.

GROSS, Gaston (1989) : *Les constructions converses du français*, Genève – Paris, Librairie Droz.

GROSS, Maurice (1981) : « Les bases empiriques de la notion de prédicat sémantique », *Langage*, n° 63, pp. 7-52.

KULAGINA, Olga et Igor MEL'ČUK (1968) : « Sovremennoe sostojanie problemy mašinnogo perevoda » [État actuel du problème de Traduction Automatique], *Problemy kibernetiki*, v. 20, pp. 297-308.

IORDANSKAJA, Lidija (1994) : « Structure communicative de l'énoncé dans la génération automatique du texte », dans cet ouvrage.

IORDANSKAJA, Lidija, KIM, Myunghee et Alain POLGUÈRE (1994) : « Some Procedural Problems in the Implementation of Lexical Functions », Leo Wanner (dir), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam, Benjamins.

LEE, Woongjae et Martha EVENS (1994) : « Generating Cohesive Text Using Lexical Functions », Leo Wanner (dir), *Lexical Functions in Lexicography and Natural Language Processing,* Amsterdam.

MEL'ČUK, Igor (1992) : « Paraphrase et lexique : la théorie Sens-Texte et le *Dictionnaire explicatif et combinatoire* », Mel'čuk *et al.* 1992, pp. 9-58.

MEL'ČUK, Igor (1994) : « Lexical Functions : a Tool for the Description of Lexical Relations in a Lexicon », Leo Wanner (dir), *Lexical Functions in Lexicography and Natural Language Processing,* Amsterdam.

MEL'ČUK, Igor, avec N. ARBATCHEWSKY-JUMARIE, L. ELNITSKY, L. IORDANSKAJA et A. LESSARD (1984) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I,* Montréal, Les Presses de l'Université de Montréal.

MEL'ČUK, Igor, avec N. ARBATCHEWSKY-JUMARIE, L. DAGENAIS, L. ELNITSKY, L. IORDANSKAJA, M.-N. LEFEBVRE et S. MANTHA (1988) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II,* Montréal, Les Presses de l'Université de Montréal.

MEL'ČUK, Igor, avec N. ARBATCHEWSKY-JUMARIE, L. IORDANSKAJA et S. MANTHA (1992) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III,* Montréal, Les Presses de l'Université de Montréal.

WANNER, Leo et John BATEMAN (1990) : « Lexical Cooccurrence Relations in Text Generation », *Proceedings of the 5th International Workshop on Natural Language Generation,* Dawson, PA.

ŽOLKOVSKIJ, Aleksandr et Igor MEL'ČUK (1967) : « O semantičeskom sinteze », *Problemy kibernetiki,* v. 19, 177-238. [Traduction française : Zholkovskij, Alexandre, et Mel'chuk, Igor (1970) : « Sur la synthèse sémantique », *T. A. Informations,* 1970, n° 2, pp. 1-85.]

13

ETAP-3 – système de traduction automatique bidirectionnel anglais-russe et russe-anglais. État actuel

Alexandre V. LAZOURSKI

Académie des sciences de Moscou, Moscou, Russie

• *Abstract* •

ETAP-3 is a bidirectional automatic translation system from English to Russian and vice versa. It has been created by a team of 10 Russian researchers headed by Academician Apresjan. ETAP-3 is based on the linguistic theory "Meaning <=> Text" proposed by Igor Mel'čuk. The translation is effected step by step: morphological analysis, syntactic parsing which results in a dependency tree representing the syntactic structure of the analysed sentence (all the correct syntactic structures can be built for an ambiguous sentence), normalization of the rough syntactic structure, transfer, re-interpretation of the initial syntactic tree in order to obtain an extended structure better adapted for the target language syntax, morphological synthesis. The ETAP-3 dictionaries contain about 13 000 entries each. ETAP-3, which is still considered as an experimental system, has already translated several articles from Russian and English scientific magazines (in all, over 5 000 sentences for each language). The rate of translation is approximately 5 seconds per sentence of about 20 words on VAX-3100 (model 30).

ETAP-3 est un système de traduction automatique développé par une équipe de 10 chercheurs russes (5 linguistes et 5 mathématiciens-programmeurs) dirigée par l'académicien Ju. Apresjan.

Ce projet a été inspiré par le modèle linguistique « Sens <=> Texte » (voir Mel'čuk 1974). Il a démarré à la fin des années 70 dans un centre de recherche relevant du Ministère électrotechnique de l'URSS. Le premier prototype (ETAP-1) réalisé sur un ordinateur français (IRIS-50) était un système de traduction automatique de textes por-

tant sur l'électrotechnique du français vers le russe. Plus tard, le français a été remplacé par l'anglais et IRIS-50 par un ordinateur russe (ES-1033). C'est ainsi que le système de traduction automatique anglais-russe (ETAP-2) a vu le jour (voir Apresjan *et al.* 1989 et 1992).

ETAP-2 fonctionnait déjà depuis un certain temps, quand toute notre équipe, à l'invitation de l'Académie des sciences, a *déménagé,* en 1985, dans l'Institut des problèmes de transfert d'information (IPPI RAN). Ce changement d'enseigne s'est répercuté également sur le projet lui-même.

Premièrement, un module assurant l'analyse de textes russes a été ajouté au système de traduction existant, qui est devenu, de ce fait, bidirectionnel.

Deuxièmement, le système a été implanté cette fois-ci sur un ordinateur américain de la série VAX, et la décision a été prise de réécrire le logiciel (qui à l'origine était en PL-1) dans un langage de programmation plus moderne et plus usité en Russie (le langage C).

Enfin, le domaine d'application a changé et les textes en informatique se sont substitués aux textes sur l'électrotechnique.

Le système avait subi beaucoup trop de modifications pour pouvoir garder le même nom. Aussi a-t-il été rebaptisé ETAP-3 (voir Apresjan *et al.* 1990).

Pour se faire une idée complète de la structure interne du système, il aurait fallu voir d'abord tous ses éléments constitutifs et étudier ensuite les mécanismes de fonctionnement des différents éléments et de leur enchaînement logique.

Seulement, l'espace qui nous est alloué ne le permet pas. Aussi, nous proposons d'examiner en même temps les composants et les algorithmes du système ETAP-3 en lui faisant traduire une courte phrase bien connue de Chomsky (*Flying planes can be dangerous*) et en suivant toutes les transformations que cette phrase subit au fur et à mesure de sa traduction vers le russe. Tous les résultats présentés sont les résultats d'un travail réel du système, sauf que les lettres russes ont été remplacées par des caractères latins.

Donc, à l'entrée nous avons la phrase anglaise :

Flying planes can be dangerous.

L'analyse et la synthèse se font en plusieurs étapes (dont chacune est effectuée par les différents modules du système, qui sont organisés de sorte que la sortie de chaque module constitue l'entrée du module qui suit).

La première étape est celle de l'analyse morphologique. À la sortie de cette étape nous allons obtenir la représentation morphologique de la phrase : une suite de lexèmes (avec tous les homonymes possibles), accompagnés de leurs caractéristiques morphologiques :

FLY1(V,ing) PLANE1(S,sg - avion)/PLANE2(S,sg - 'surface plane')
CAN1(S,sg - 'boîte de conserve')/CAN2(V,mf - 'pouvoir')/CAN3(V,mf-
'mettre en boîte') BE(V,inf) DANGEROUS(A)

En même temps, le système accède aux informations, très nombreuses et diversifiées, qui sont inscrites dans les articles du dictionnaire, dit *combinatoire.* Par souci d'économie, je ne décrirai pas en détail les entrées du dictionnaire (une entrée peut contenir plus de 100 lignes d'informations). Il suffira de dire tout simplement que chacune des entrées présente une description aussi complète que possible des caractéristiques syntaxiques et sémantiques du mot (qui déterminent son aptitude ou son incapacité de faire partie de telle ou telle construction syntaxique), et contient de nombreuses références aux règles spécifiques de différents niveaux qui sont significatives pour ce mot.

À partir de la représentation morphologique de la phrase on peut commencer l'analyse syntaxique. Cette deuxième grande étape est précédée, cependant, par une procédure qui s'appelle *analyse présyntaxique* et dont le but est de résoudre les cas d'homonymie en se basant uniquement sur le contexte linéaire du mot homonymique. Ainsi, dans notre exemple l'analyse présyntaxique éliminera certaines hypothèses :

FLY1(V,ing) PLANE1(S,sg - avion)/PLANE2(S,sg - 'surface plane')
CAN2(V,mf - 'pouvoir') BE(V,inf) DANGEROUS(A)

L'objectif de l'analyse syntaxique est de construire pour la phrase analysée un graphe orienté, dont les nœuds seront représentés par les mots de la phrase, et les branches par les relations syntaxiques (ETAP-3 en distingue une cinquantaine pour chacune des langues) reliant les mots à l'intérieur de cette phrase. Cet arbre de dépendances est construit à l'aide des règles syntaxiques, dites *syntagmes.*

Dans un premier temps, le système scrute les syntagmes et retient les hypothèses qui correspondent aux caractéristiques morpho-syntaxiques des membres de la phrase analysée :

1.1 FLY1	-->	2.2 PLANE1	- 1-COMPL.01.1
			- 1-COMPL.12.1
2.1 PLANE2	-->	1.1 FLY1	- MODIF.01.1
2.2 PLANE1	-->	1.1 FLY1	- MODIF.01.1
2.1 PLANE2	-->	5.1 DANGEROUS	- MODIF.02.1
2.2 PLANE1	-->	5.1 DANGEROUS	- MODIF.02.1
3.2 CAN2	-->	1.1 FLY1	- PREDIC.05.1
			- ADVERB.02.1
			- ADVERB.04.1
3.2 CAN2	-->	2.1 PLANE2	- PREDIC.01.1
3.2 CAN2	-->	2.2 PLANE1	- PREDIC.01.1
3.2 CAN2	-->	4.1 BE	- 1-COMPL.17.1
4.1 BE	-->	5.1 DANGEROUS	- COPULAT.10.2

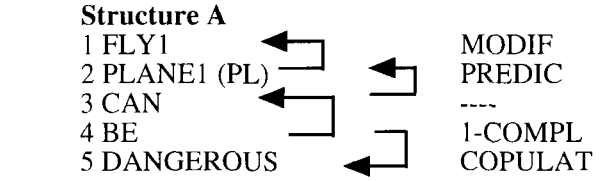
Ensuite, il essaie d'établir le sommet syntaxique de la phrase. Dans notre cas de figure c'est très simple, parce que pour le mot numéro 3 de la phrase (CAN2) il n'y a pas une seule hypothèse suivant laquelle il puisse occuper une position syntaxique

subordonnée. Par conséquent, ce mot *maître par excellence* ne peut être que sommet de la phrase[1].

Puis, le système identifie tous les *liens uniques* (pour lesquels il n'existe pas d'alternative). Dans notre exemple, c'est le cas de la relation syntaxique complétive entre les mots CAN2 et BE (le mot CAN2 est le seul mot de la phrase qui prétend assujettir le mot BE). Ensuite, le système examine une fois de plus les syntagmes qui correspondent *aux liens uniques* pour éliminer toutes les hypothèses qui ne sont pas compatibles avec ces derniers. C'est un processus itératif, parce que l'élimination de certaines hypothèses peut conduire à la création de nouveaux *liens uniques*. Mais, arrive le moment où la situation se stabilise et l'on ne peut plus trouver de contradictions qui permettraient de résoudre l'homonymie syntaxique en supprimant les hypothèses notoirement fausses :

1.1 FLY1	-->	.2 PLANE1	- 1-COMPL.01.1
2.1 PLANE2	-->	.1 FLY1	- MODIF.01.1
2.2 PLANE1	-->	.1 FLY1	- MODIF.01.1
3.2 CAN	-->	.1 FLY1	- PREDIC.05.1
3.2 CAN	-->	.1 PLANE2	- PREDIC.01.1
3.2 CAN	-->	.2 PLANE1	- PREDIC.01.1
3.2 CAN	-->	.1 BE	- 1-COMPL.17.1
4.1 BE	-->	.1 DANGEROUS	- COPULAT.10.2

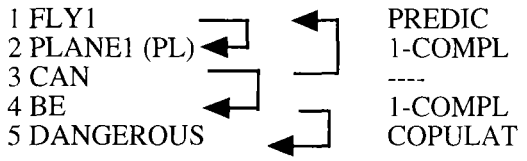
Alors, le système commence à envisager les éventualités : pour le mot à plusieurs maîtres syntaxiques possibles, il retient une hypothèse et considère toutes les conséquences qui en découlent. Il se peut que cette hypothèse soit fausse, auquel cas le système (tôt ou tard) se heurte à une anomalie et revient en arrière pour envisager d'autres hypothèses. Mais si l'hypothèse est fondée, le système finira par construire un arbre complet qui lui correspond :



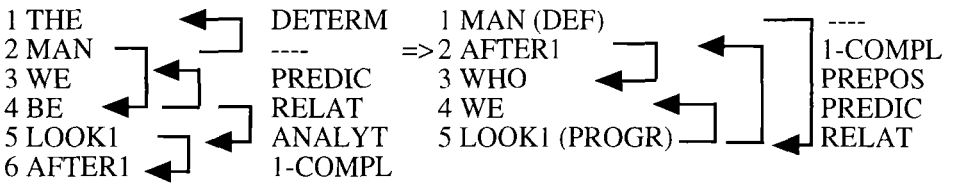
On peut également lui demander de rechercher toutes les autres hypothèses (provisoirement abandonnées) qui aboutissent à des structures syntaxiques complètes non contradictoires :

1. C'est d'ailleurs assez rare que le sommet de la phrase soit identifié ainsi : d'entrée de jeu, et sans ambiguïté possible. Normalement, au début de l'analyse, chacun des mots de la phrase a au moins un lien syntaxique hypothétique qui le lie à un autre mot. À ce moment-là, comme le critère du *maître par excellence* ne s'applique pas, le système applique une règle spéciale qui reconnaît tous les sommets potentiels de la phrase, et leur attribue des priorités en se basant sur de nombreux critères. Par la suite, toutes les possibilités sont envisagées l'une après l'autre suivant les priorités établies.

Structure B

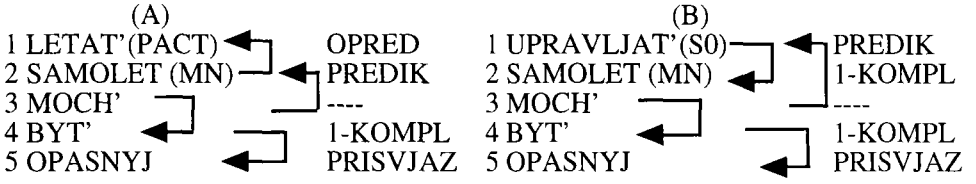


Le stade suivant est la normalisation de la structure syntaxique obtenue, c'est-à-dire l'élimination de certains traits excessivement idiomatiques et trop colorés de la syntaxe de la langue de départ. Par exemple, à l'étape de la normalisation, la construction syntaxique typiquement anglaise, à préposition non saturée dans la subordonnée relative : *The man we are looking after* sera réduite à une expression beaucoup moins idiomatique mais plus explicite : *The man after whom we are looking.* En même temps, l'article sera effacé (parce que la catégorie grammaticale qui lui correspond s'exprime différemment d'une langue à l'autre), et sa signification sera transformée en une caractéristique grammaticale universelle – DEF(initif). Il en va de même pour l'auxiliaire de la forme analytique du *present continuous*, qui sera supprimé, alors que la caractéristique PROGR(essif) sera attribuée au verbe principal :



Pour la phrase de Chomsky, l'action du module de normalisation ne sera pas aussi flagrante et spectaculaire et se limitera uniquement à l'explication de l'ambiguïté syntaxique de la *ing-form* qui, en anglais, peut remplir trois rôles syntaxiques différents : celui d'un adjectif (*a sleeping beauty*), d'un substantif (*the thinking aloud*) et d'un adverbe (*he sat doing nothing*). Il faut noter que cette polyvalence syntaxique du participe présent est propre à l'anglais, mais n'existe pas dans d'autres langues. Aussi la tâche du module de normalisation pour les *ing-forms* sera-t-elle d'expliciter leur rôle syntaxique en leur rattachant une des trois caractéristiques : *PACT, S0* ou *GER*, qui renvoient à des catégories grammaticales plus universelles. Dans les exemples considérés le mot *flying* sera interprété comme PACT dans la structure A et comme S0 dans la structure B. À une des étapes ultérieures ces caractéristiques donneront lieu à la génération en russe d'un participe présent (pour le PACT) et d'un substantif dérivé du verbe (pour le S0).

Ensuite vient le moment du transfert. Par la phrase analysée (qui ne contient pas de terminologie spéciale ou de tournures plus ou moins idiomatiques), le transfert ne présente aucune particularité et sera réduit à sa plus simple expression : les mots anglais seront remplacés par des mots russes (inscrits dans la zone de traduction des entrées correspondantes) ; les relations syntaxiques anglaises de la structure normalisée seront transformées en relations syntaxiques russes, et les caractéristiques morphologiques significatives anglaises en caractéristiques morphologiques russes :



L'étape suivante (extension) est très similaire à celle de la normalisation, mais prise à l'envers. Si le module de normalisation doit porter les traits syntaxiques trop idiomatiques de la langue de départ à un certain point de simplification et d'unification, le module d'extension, au contraire, vise à générer les traits lexico-syntaxiques idiomatiques, propres à la langue d'arrivée, à partir d'une structure syntaxique primaire. C'est à cette étape-là que se fera l'interprétation des caractéristiques syntaxique : PACT engendrera un participe présent, actif, non perfectif du verbe *letat'* et S0 donnera lieu à une dérivation lexicale, le verbe *upravljat'* sera remplacé par le substantif correspondant (*upravlenie*).

Le module suivant (celui de la synthèse syntaxique) sert à générer les caractéristiques morphologiques chez les membres de la phrase, en fonction des liens syntaxiques qui les relient les uns aux autres.

Enfin le dernier module destiné à la synthèse morphologique sert à générer les formes de mots correctes, conformes aux règles d'orthographe de la langue d'arrivée, à partir des noms de lexèmes, munis à l'étape précédente de toutes les caractéristiques morphologiques. À la sortie de ce module on obtient la traduction finale de la phrase analysée :

(A) *Letajushchie samolety mogut byt' opasnymi.*
(B) *Upravlenie samoletami mozhet byt' opasnym.*

La série ETAP a été conçue comme entièrement automatique, c'est-à-dire, qu'elle évite toute intervention humaine en cours de traduction. Ceci étant, la dernière version possède plusieurs modes de fonctionnement différents qui offrent aux utilisateurs des possibilités très variées.

Le premier mode est complètement autonome : le texte de départ, enregistré sur un support magnétique, apparaît dans la partie supérieure de l'écran (la phrase analysée étant mise en surbrillance), alors que, dans la partie inférieure, commencent à défiler les phrases traduites dans la langue-cible.

Le deuxième mode est un système d'aide au traducteur. Il est accessible à partir d'un éditeur de textes. Le traducteur humain fait défiler le texte dans une fenêtre et marque les fragments qu'il veut faire traduire par l'ordinateur (ça peut être le texte tout entier, un alinéa, une phrase, une partie de la phrase, ou encore un mot ou une expression terminologique qu'il ne connaît pas). Les résultats de la traduction vont apparaître dans la deuxième fenêtre. Si le texte analysé présente une homonymie syntaxique ou lexicale quelconque, on peut exiger que toutes les variantes de traduction possibles soient affichées.

Enfin, le troisième mode correspond plutôt aux besoins des linguistes auteurs du

système. Sous ce mode-là, les protocoles de travail détaillés sont sauvegardés, ce qui permet de vérifier pas à pas le fonctionnement du système, de localiser les anomalies et les erreurs et d'apporter les corrections nécessaires, puisque tous les éléments constitutifs du système (dictionnaires, listes des caractéristiques, règles, etc.) sont directement accessibles. En fait, c'est un véritable poste de LAO (linguistique assisté par ordinateur). Initialement destiné à la mise au point des données linguistiques, des algorithmes et des logiciels, il ouvre aujourd'hui de nouvelles perspectives, parce que, grâce aux développements tout récents, il permettra bientôt d'automatiser une grande partie des travaux de routine, liés à la rédaction des dictionnaires du système. Et là, on touche, vraiment, au point sensible.

Bien que ETAP-3 soit basé sur un modèle très élaboré de morphologie et de syntaxe des deux langues (anglais et russe), que l'on croit suffisamment complet pour assurer l'analyse et la synthèse automatiques de textes dans divers domaines scientifiques et techniques, une application industrielle du système n'est pas possible dans l'immédiat à cause de la taille des dictionnaires (les dictionnaires du système à l'état actuel se chiffrent à quelque 12 milliers d'entrées pour chacune des langues).

À ce jour le système a traduit plusieurs milliers de phrases de l'anglais vers le russe et du russe vers l'anglais. Pendant tout ce temps-là, nous n'avons pas cessé d'apporter des corrections et des modifications au système pour tenir compte des erreurs détectées au cours des essais. Les résultats des derniers tests-témoins ont montré une tendance à la stabilisation du système. Quant à la rapidité du système, le temps moyen de traduction d'une phrase d'une vingtaine de mots sur le microVAX-3100 (modèle 30) est de l'ordre de cinq secondes.

En conclusion, j'aimerais bien mettre en relief quelques traits conceptuels propres à l'ETAP-3, qui se traduisent en avantages pratiques non négligeables.

- Les langues opérationnelles sont décrites d'une façon tout à fait indépendante l'une de l'autre. Autrement dit, le système ignore quelle sera la langue-cible jusqu'au moment où il arrive à l'étape du transfert à proprement parler. Cette indépendance présente deux avantages importants : a) les mêmes données linguistiques peuvent être utilisées pour l'analyse et pour la synthèse (avantage dont nous avons largement profité en passant du système de traduction unidirectionnel anglais-russe ETAP-2 au système bidirectionnel) ; b) lorsqu'une nouvelle langue est intégrée au système, il suffit seulement de décrire cette langue, sans qu'il soit nécessaire de changer la description des autres langues faisant déjà partie du système.

- Les données linguistiques sont présentées sous une forme déclarative (donc indépendante des algorithmes du système) et suivant les modèles standardisés, dits *formats de description.* Les algorithmes sont ajustés aux formats et font complètement abstraction de leur contenu. Par conséquent l'ajout d'une nouvelle langue n'implique aucune modification au niveau des algorithmes.

Références

APRESJAN, Ju. D., BOGUSLAVSKY, I. M., IOMDIN, L. L., LAZURSKIJ, A. V., PERTSOV, N. V., SANNIKOV, V. Z. et L. L. TSINMAN (1989) : *Ling-*

visticeskoe obespecenie sistemy ETAP-2, [Partie linguistique du système ETAP-2], Moscou, Nauka, 295 p. (En russe).

APRESJAN, Ju. D, BOGUSLAVSKY, I. M., IOMDIN, L. L. *et al.* (1990) : *A Bi-directional Machine Translation System ETAP-3: Current State,* ML-ETH Prague-4 Report, European Coordination Centre for Research and Documentation in Social Sciences, Vienne.

APRESJAN, Ju. D., BOGUSLAVSKY, I. M., IOMDIN, L. L., LAZURSKIJ, A. V., PERTSOV, N. V., SANNIKOV, V. Z. et L. L. TSINMAN (1992) : « ETAP-2: The Linguistics of a Machine Translation System », *Meta,* vol. 32, n° 1, pp. 97-112.

MEL'ČUK, I. A. (1974) : *Opyt teorii lingvisticeskix modelej « Smysl <=> Tekst »*, [Essai d'une théorie de modèles linguistiques « Sens <=> Texte »], Moscou, Nauka, 314 p. (En russe).

14

Pour des systèmes de TA adaptifs multi-architectures

Sergei NIRENBURG*, David FARWELL** et Yorick WILKS**[1]

• *Abstract* •

A number of proposals have come up in recent years for hybridization of MT. Current MT projects – both "pure" and hybrid, both predominantly technology-oriented and scientific (including those currently funded by NSF) are single-engine projects, capable of one particular type of source text analysis, one particular method of finding target language correspondences for source language elements and one prescribed method of generating the target text. While such projects can be quite useful, we believed that it is time to make the next step in the design of machine translation systems and to move toward adaptive, multiple-engine systems. We describe the architecture of an adaptive multi-engine MT system which uses each of the engines under the circumstances which are most favorable for its success.

Pour des systèmes de TA adaptifs multi-architectures

Ces dernières années ont été le témoin d'une modification de l'équilibre entre les efforts technologiques et scientifiques dans le domaine de la TA. Les dernières nouveautés méthodologiques du domaine sont essentiellement technologiques et n'ont pas pour but de faire avancer notre connaissance sur les mécanismes de base de compréhension et de production de textes ou sur les modèles informatiques qui simulent ces mécanismes.

Les deux paradigmes les plus en vogue en TA – la traduction basée sur les exem-

*Center for Machine Translation, Université Carnegie Mellon, Pittsburgh, États-Unis.
**Computing Research Laboratory, Université New Mexico State, Las Cruces, États-Unis.
1. Les auteurs remercient Ted Dunning pour ses interventions.

ples (TABE) et la traduction basée sur les statistiques (TABS) – ne requièrent qu'une connaissance *a posteriori* du langage. Alors que les exemples représentatifs de ces paradigmes en sont encore au stade de systèmes pilotes (par exemple, Furuse et Iida 1992 ; McLean 1992 ; Jones 1992 et Maruyama et Watanabe 1992) et se débattent avec les contraintes inhérentes aux approches qui évitent l'étude du langage comme tel (par exemple, Brown *et al.* 1990), différentes propositions ont été émises pour une hybridation de la TA. Dans certaines de ces approches, l'analyse du corpus est utilisée pour la mise au point de l'analyse et des grammaires de transfert (par exemple, Su et Chang 1992) ; dans d'autres, l'approche standard de transfert (TABT) utilise les techniques traditionnelles d'analyse et de génération, mais s'appuie sur une composante de transfert basée sur des corpus bilingues alignés (par exemple, Grishman et Kosaka 1992) ; dans d'autres encore, il est suggéré que les informations statistiques soient utilisées comme source d'assignation de préférence durant la désambiguïsation du texte (voir par exemple, le résumé présenté dans Lehmann et Ott 1992). Les systèmes de TA hybrides furent d'ailleurs au centre des débats du dernier colloque international sur les aspects théoriques et méthodologiques de la traduction automatique.

Il est cependant important de reconnaître que la plupart des propositions en faveur de l'hybridation reposent essentiellement sur des considérations technologiques. Bien sûr, la TA est un domaine de recherche appliquée et il est parfaitement normal que l'incitant pour le progrès vienne, pour une large part, de sources extra-scientifiques. Toutefois, la TA est une application particulière. Contrairement à de nombreux autres domaines, c'est un excellent champ d'expérimentation pour les théories linguistiques (syntaxiques, sémantiques, pragmatiques ou discursives), pour les méthodes de linguistique computationnelle (algorithme de parsage, interprétation sémantique et pragmatique et génération de texte), pour la linguistique descriptive (lexiques et grammaires pour des langages particuliers), pour les processus de modélisation du raisonnement humain (représentation des connaissances et leur manipulation) et pour les études de traduction. En fait, il s'agit pour nous d'un environnement idéal pour développer et tester des traitements auxquels nous avons fait référence sous le terme de « microthéories » (Nirenburg *et al.* 1992), c'est-à-dire des traitements d'un large éventail de phénomènes spécifiques du langage tels que les dépendances sémantiques tête-modifieur, l'aspect ou la quantification, ainsi que des théories computationnelles plus complètes de l'usage de la langue.

Les projets récents en TA – tant « purs » qu'hybrides et tant ceux à prédominance scientifique que ceux orientés technologie, sont des projets mono-architecture, capables d'un seul type d'analyse du texte source et disposant d'une seule méthode pour trouver les correspondances cibles des éléments du langage source et pour la génération du texte du langage cible. Quoique de tels projets puissent être utiles, il est temps, selon nous, d'aller plus loin dans la conception de systèmes de TA et d'évoluer vers des systèmes multi-architectures adaptifs.

De plus, les systèmes courants ont été développés pour un type de texte particulier (bulletins météorologiques, articles d'actualités financières, résumés scientifiques) et pour un usage déterminé (assimilation ou diffusion de l'information). Étant donné la spécification du type de texte en entrée et son utilisation finale, l'un des systèmes sera le plus approprié. Ainsi, par exemple, pour traduire en masse des résumés scientifiques d'un domaine particulier dans le but d'informer une audience particulière du contenu des derniers articles dans le domaine, la TA basée sur les exem-

ples semble être préférable. Pour traiter de courts articles concernant un large éventail de sujets dans le but de sélectionner ceux qui apparaissent d'un intérêt particulier, une approche basée sur les statistiques semble la plus appropriée. Cette correspondance entre technique, type de texte source et utilisation finale (ou type de texte cible) plaide pour des systèmes adaptifs multi-architectures.

Dans le camp des chercheurs en TA rationalistes, la question du niveau de complexité de l'analyse du texte source a été, pendant longtemps, au centre du débat scientifique. Les lignes de démarcation étaient tracées comme suit. Un groupe de chercheurs prétendait que, sauf si un nombre étendu de phénomènes linguistiques qui apparaissent dans les textes naturels est analysé et largement représenté, la traduction de qualité est impossible. Les adhérents de cette approche orientée-signification affirment que de tels phénomènes ne peuvent être décrits qu'avec l'aide d'abondantes connaissances stockées dans des grammaires et des lexiques (généralement traités manuellement) détaillés de chaque langue concernée.

Un autre groupe de chercheurs prétend que cette tâche d'acquisition des connaissances n'est pas réaliste. En se basant sur de nombreuses observations qui montrent qu'une analyse profonde n'est pas toujours nécessaire pour la traduction (par exemple, le terme polysémique espagnol *centro* sera traduit en allemand *zentrum* quel que soit le sens de *centro* dans le texte source), ces chercheurs optent pour une analyse plus simple et l'utilisation de substitutions plus directes en langage cible plutôt qu'une analyse de la signification concernée.

Une formulation caractéristique de cette position est donnée par Ben Ari *et al.* (1988 : 2) :

> « il faut bien garder à l'esprit que le processus de traduction ne nécessite pas une compréhension complète du texte. Beaucoup d'ambiguïtés peuvent être conservées durant la traduction..., et ne devraient pas être présentées à l'utilisateur (traducteur) pour résolution ».

De même, Isabelle et Bourdeau (1985 : 21) affirment que :

> « quelquefois, il est possible d'ignorer certaines ambiguïtés dans l'espoir que ces mêmes ambiguïtés seront conservées lors de la traduction. C'est particulièrement vrai pour des systèmes comme TAUM-AVIATION qui traitent une seule paire de langues étroitement liées pour un sous-langage restreint. La problématique de l'attachement de la phrase prépositionnelle, par exemple, est fréquemment contournée par ce biais. De manière générale, toutefois, l'analyse tend à produire une représentation intermédiaire non ambiguë. »

Notre approche provient de la volonté d'être en mesure de ne réaliser que le travail absolument nécessaire dans le processus de traduction, en optimisant l'ensemble du système par les types de travaux que chaque module réalise le mieux. Ainsi, si un passage source est identique (ou très similaire) à un passage antérieurement traduit et stocké dans une base de données de traductions anciennes, l'utilisation d'une architecture de TABE qui n'implique que la recherche de traductions mémorisées, sera indiquée. Mettre en route l'ensemble du mécanisme de, admettons, la TABC sera, dans ce cas, une perte de ressources.

Quoique nous poursuivions nos recherches sur l'acquisition (semi-)automatique des connaissances (Wilks *et al.* 1990), nous croyons également qu'un système de TA adaptif qui comporterait, en complément de composantes informatico-linguistiques de haut niveau, une composante statistique et une composante basée sur les exemples, peut, dans bien des circonstances, dégager le goulet d'étranglement des connaissances. La qualité et la rapidité du matériel actuel rend possible un système composé de différentes architectures de TA.

Un système adaptif, qui permettrait des substitutions directes quand elles sont possibles, mais activerait des architectures avec des composantes d'analyse progressivement plus complexes quand c'est nécessaire, est une solution préférable au double danger, d'une part, d'économiser les ressources et du temps-machine tout en courant un risque d'erreurs dues au manque de connaissances décisionnelles dans un système de haut-niveau, ou d'autre part, de construire un système qui est *a priori* incapable de traiter un large champ de la signification du langage naturel.

Les modules

Le module de base de TABC de notre système sera l'architecture de TA Pangloss. Le passage source est soumis à l'analyse morphologique et syntaxique qui produit une structure syntaxique. Le résultat est soumis à un interpréteur sémantique, qui produit un ensemble de structures sémantiques et un ensemble d'unités de connaissance décrivant la situation. Ce résultat est soumis à un générateur du langage cible qui délimite les phrases cibles et les organise ensuite sur les plans lexical, syntaxique et morphologique. Le processus global est supporté par des sources de connaissances statiques, comme les grammaires et les lexiques source et cible. Dans le lexique, les champs sémantiques des entrées pour les items lexicaux de la classe ouverte contiennent une explication de leurs significations par une référence à un concept ou à un groupe de concepts dans un modèle formel du domaine, une ontologie. L'analyseur sémantique dispose parfois de connaissances insuffisantes pour réaliser une assignation particulière ou une désambiguïsation. Dans ce cas, une interface interactive, l'augmenteur, (voir, pour une description plus détaillée, Nirenburg *et al.* 1992 ; Brown et Nirenburg 1990 ; Brown 1990) est utilisée pour permettre à un opérateur humain d'aider le système à prendre la décision correspondante. Cette interface sera particulièrement efficiente pour le traitement de segments sources inattendus (mots inconnus, noms propres, fautes de frappe, etc.). Le progrès dans le développement d'un tel système de TABC peut être mesuré par le taux décroissant des questions à l'augmenteur.

Le module TABT utilisera l'analyseur syntaxique de Pangloss et remplacera l'interpréteur sémantique par une composante de transfert lexical et structural. Il produira, en résultat, des structures dans un format utilisable aussi bien par le générateur Penman que par DIOGENES (Nirenburg *et al.* 1992) – le choix sera fait ultérieurement. Le but principal de la composante de transfert sera de remplacer les éléments ontologiques qui représentent les items lexicaux sources dans le formalisme de représentation par des items lexicaux cibles. L'ensemble de règles de transfert syntaxique se concentrera sur le traitement des lexèmes avec des arguments, en réalisant des mises en correspondances directes entre des schémas de sous-catégorisation de lexèmes correspondants source et cible.

Le module de TABE sera organisé comme suit. Un passage source sera comparé au côté source d'un corpus bilingue existant. Si la mise en correspondance réussit, le côté cible du passage sélectionné dans le corpus sera pris comme résultat de la traduction. Dans cette approche, la tâche centrale est le développement d'une mesure efficace de la « distance » qui existe entre deux passages. La mesure la plus simple est une correspondance totale. Toutefois, dans la pratique, cette situation ne se produit pas très souvent et il faut définir des mesures plus complexes. À ce jour, les expérimentations de TABE (Furuse et Iida 1992) utilisent des thesaurus et des dictionnaires *on-line* pour juger de la « proximité » de certains lexèmes (par exemple, *maison* sera plus proche de *construction* que de *chambre).* Pour la mesure de la distance, nous proposons d'utiliser, en plus d'un dictionnaire lisible sur machine, l'ontologie que nous avons développée pour l'architecture de TABC. Ce sera ainsi le premier système de TABE construit sur cette base, avec, comme résultat, une mise en correspondance plus efficiente.

Nous nous proposons d'importer entièrement le module de TABS. Les ressources seront ainsi entièrement consacrées à l'intégration d'une structure multi-architectures adaptive.

Le répartisseur

Dans notre système, les passages sources sont assignés aux différentes architectures de TA par le répartisseur. Sa fonction est de segmenter le texte source et d'assigner chaque segment à l'une des quatre architectures de TA : TABS, TABE, TABT, TABC. Il prend pour entrée le texte (et certaines de ses spécifications générales) et réalise un ensemble de diagnostics pour établir les unités de traduction et sélectionner la meilleure architecture. Il distribue ensuite les segments de texte à la meilleure architecture. Le répartisseur évaluera chaque passage en fonction des critères généraux suivants :

- Le type de traduction – Le résultat de la traduction a-t-il pour but la diffusion (c'est-à-dire la lecture à l'extérieur d'une organisation) ou l'assimilation (c'est-à-dire, la lecture à l'intérieur d'une organisation) ? Une traduction complète est-elle nécessaire ou peut-on se contenter d'un résumé, ou bien encore, d'une simple catégorisation du texte (le texte est-il suffisamment important pour être traduit dans sa totalité ?) ? Les systèmes de TABC et de TABT devraient produire les traductions de meilleure qualité, sauf si l'*input* correspond entièrement à un passage d'un corpus aligné. Les traductions destinées à l'assimilation peuvent être moins « peaufinées » et certaines erreurs et omissions peuvent être tolérées, ce qui peut faire pencher la décision au profit d'architectures moins sophistiquées, telles les architectures de TABE ou de TABS.

- L'existence de textes parallèles dans un domaine ou sur un sujet particulier, une condition importante pour le choix d'un système de TABE ou de TABS. La qualité d'un *output* de TABS ne peut être déterminée avant d'avoir traduit entièrement le passage (Peter Brown, communication personnelle). Le module de TABS est le plus expérimental de toutes les architectures de notre système adaptif. Tout comme pour les autres modules,

les décisions du répartisseur concernant cette architecture seront dérivées empiriquement pendant l'évaluation et la phase de test de notre système.

- Le degré d'ambiguïté d'un passage source, tant dans la langue source elle-même que vis-à-vis de la langue cible (y compris le phénomène d'imprécision (*vagueness*)). Dans le premier cas, la règle générale est que plus le degré d'ambiguïté est élevé, plus l'approche de TABC est indiquée. Toutefois, si le langage cible traduit de manière identique les ambiguïtés du langage source (rappelons l'exemple *centro* —> *zentrum*), l'approche du transfert offre une solution plus simple. En cas d'imprécision, des solutions pragmatiques imposeront le choix de l'architecture de TABC. S'il est possible de trouver des critères de solution plus simples pour des cas particuliers d'imprécision (comme, par exemple, pour les collocations), l'approche de transfert peut être réutilisée à nouveau.

- La taille et la qualité des ressources de TABC disponibles (ontologie, lexique, etc.). Si l'on détermine que, pour un passage source donné, la couverture des connaissances statiques disponibles est sérieusement incomplète, on peut prévoir que la qualité du résultat sera inférieure. Le répartisseur pourra juger que la quantité d'interactions avec l'augmenteur, requise pour corriger la situation, est trop élevée et favoriser dans ce cas une architecture de TABE, même si la qualité du résultat devait en souffrir.

Le rôle du répartisseur inclut donc la construction de programmes pour : a) évaluer le contexte de traduction conformément aux quatre critères énoncés ci-dessus et b) disposer d'un mécanisme de décision pour calculer le degré d'adéquation de chacune des architectures pour le traitement d'un texte source dans un contexte donné.

Un paramètre additionnel important dans l'opération du répartisseur est de déterminer la taille la plus appropriée d'un passage source. Puisqu'un texte source peut être traité par une combinaison d'architectures de TA, la qualité attendue du résultat dépendra des différentes manières possibles de segmenter le texte source. Le répartisseur sera donc chargé de trouver la longueur optimale de chaque passage devant être assigné à une architecture particulière.

Le répartisseur utilisera un ensemble additionnel de diagnostics déterminés par la structure de l'architecture spécifique de TA. Le développement de ces heuristiques est un des points centraux de notre recherche. Nous proposerons dans les paragraphes suivants une analyse préliminaire de ces heuristiques de diagnostics spécifiques, ordonnées par type d'architecture.

TABS Nous préconisons une approche basée sur les statistiques qui utilise un modèle statistique composé d'un modèle de traduction des mots LS-LC, d'un modèle d'alignement LS-LC et d'un modèle d'ordonnancement LC. Le processus consiste à appliquer l'analyseur morphologique pour convertir la chaîne d'entrée en une chaîne de pseudoformes avec des informations catégorielles et flexionnelles, à réordonner de manière sélective les éléments, pour diviser ces chaînes « pseudoformes » en sous-chaînes traitables, à appliquer le modèle statistique composite LS-LC, à induire une sous-chaîne de « pseudoformes » équivalente en LC et à appliquer le générateur mor-

phologique LC pour produire l'expression dans le LC. Le résultat est ensuite post-édité. Les ressources de la TABS incluent :

- un analyseur morphologique qui produit des pseudoformes avec des informations catégorielles et flexionnelles ;
- un jeu de règles de réordonnancement pour juxtaposer les têtes des constructions lexicales ;
- un segmenteur de sous-chaînes qui divise l'entrée en sous-chaînes de longueur appropriée ;
- une base de données avec les correspondances LS-LC ;
- une base de données avec les correspondances d'alignements LS-LC ;
- une base de données des trigrammes LC pour contraindre les ordonnancements cibles ;
- un générateur morphologique du LC qui produit des formes fléchies pour des pseudoformes, en utilisant les informations catégorielles et flexionnelles.

Les diagnostics heuristiques possibles sont les suivants :

- analyse de la fréquence d'occurrence de chaque item individuel dans le corpus. Plus grande est la fréquence des items contenus dans le texte, plus grande est la probabilité que la base empirique pour les statistiques, utilisée pour conduire le processus, soit adéquate, et donc, plus grande est la probabilité que l'architecture TABS produise le résultat désiré ;

- comparaison de chaque item ou séquence contiguë d'items de la chaîne *input* avec une liste de formes de citations connues. Plus élevé est le nombre de formes « inconnues » (formes pour lesquelles il n'y a pas d'équivalence), moins il y a de probabilité que l'architecture TABS produise une traduction adéquate ;

- calcul du degré de confiance associé avec la division de la chaîne d'entrée en sous-chaînes ; plus le degré de confiance est faible, moins le système TABS produira les résultats désirés.

Pour chacun de ces diagnostics, des tests aideront à établir les seuils critiques.

TABE La structure du module de TABE a été esquissée ci-dessus. Les ressources du mécanisme de traduction TABE comprennent :

- un segmenteur de phrases qui produit une segmentation au niveau de la phrase du texte en entrée ;
- un analyseur morphologique qui produit nœuds, catégories et informations flexionnelles pour les divers items de la chaîne d'entrée ;
- une base de données d'équivalents lexicaux et phrasaux LS-LC ;
- une procédure de comparaison des schémas LS, qui incorpore une métrique de similarité.

Pour la TABE, il est plus facile de développer des diagnostics puisque c'est essentiellement une opération de comparaison de schémas. Les indicateurs potentiels de la capacité de l'architecture TABE à produire une traduction adéquate sont les suivants :

- calcul pour chacun des items de la chaîne d'entrée de leur fréquence d'occurrence dans les corpus. Plus grandes sont les fréquences des items du texte en entrée, plus grande est la probabilité que des exemples significatifs soient contenus dans la base de données des équivalents LS-LC (voir, architecture TABS) ;

- vérification, dans la chaîne d'entrée, de la présence de formes inconnues (formes qui ne sont pas dans la base de données avec les équivalents LS-LC). À nouveau, plus grand est le nombre de formes « inconnues » dans l'entrée, plus faible est la probabilité que l'architecture TABE produise des résultats adéquats (identique à l'architecture TABS) ;

- sans appliquer la métrique de similarité, comparaison des séquences des items dans la chaîne en entrée avec les exemples dans la base de données avec les équivalents LS-LC. Plus grand est le nombre de chaînes *input* qui sont tout à fait similaires à celles de la base de données, plus grande est la probabilité de succès d'application d'une architecture TABE. Les sous-segments du texte *input* peuvent aussi bénéficier d'un facteur de confiance plus grand ou plus faible en fonction du degré avec lequel ils sont complètement couverts par le processus de mise en correspondance.

TABT Pour l'architecture TABT, nous préconisons une architecture avec un transfert au niveau de la structure de la phrase et qui implique la substitution des arbres et des nœuds lexicaux. Le processus de TABT a été esquissé ci-dessus. Les ressources du mécanisme comportent :

- un segmenteur de phrases divisant le flux du texte en entrée en phrases ;
- un analyseur morphologique convertissant les mots en nœuds marqués de la catégorie et d'informations flexionnelles ;
- un analyseur de structure de phrase produisant les diagrammes de dépendance syntaxique ;
- un composant de transfert arbre à arbre, basé sur des informations contrastives, qui substitue les items du LS en items du LC avec les modifications structurelles correspondantes ;
- un inflecteur TL produisant des expressions du LC pour les nœuds.

Les diagnostics sont les suivants :

- la comparaison de chaque item de la chaîne en entrée avec les items lexicaux du système d'analyse du langage source. Moins il y a de formes « inconnues », plus la probabilité de succès est grande. Néanmoins, plus grand est le degré d'ambiguïté (c'est-à-dire plus grand est le nombre d'items du LS produisant de nombreux équivalents en LC), moins il y a de probabilité de succès ;

- le marquage de la chaîne d'entrée avec les « parties du discours » ; comparaison des résultats avec les côtés droits des règles de grammaire en LS. Plus le degré de couverture est élevé, plus grande est la probabilité de succès. Mais, à nouveau, plus grand est le degré d'ambiguïté de la couverture (c'est-à-dire les séquences « partie du discours » auxquelles correspondent plusieurs

règles PS), moins élevée est la probabilité de réussite. Si la composante de transfert devait se baser sur des séquences « partie du discours » LS, ces règles pourraient également être utilisées comme base de calcul de la couverture.

TABC Enfin, nous préconisons pour l'architecture TABC une approche basée sur les connaissances susceptibles d'utiliser le système Pangloss. Le processus a été brièvement esquissé ci-dessus. Les ressources du mécanisme impliquent :

- un segmenteur de phrases qui divise le flux de texte source en phrases ;
- un analyseur morphologique qui convertit les mots en nœuds marqués par des traits catégoriels et flexionnels ;
- un module qui fournit les segments niveau phrase (*chunker*) ;
- un parseur niveau phrase, qui produit une représentation explicite des informations morphologiques, syntaxiques (internes et niveau phrase) et pragmatiques des morceaux de phrase ;
- un assembleur niveau clause qui construit la structure sémantique ;
- un module qui fait le lien entre la représentation sémantique et la représentation du sens du texte et qui est responsable des informations sémantiques et pragmatiques au niveau de la phrase ;
- un planificateur de phrases LC qui arrange les clauses de représentation du sens du texte en des séquences de phrases complexes ;
- un générateur LC qui sélectionne les items lexicaux appropriés et fournit les diagrammes de dépendance syntaxique pour les phrases ;
- un inflecteur LC qui fournit les formes graphiques pour les nœuds des diagrammes de dépendance syntaxique.

Les diagnostics pour l'architecture TABC recouvrent en quelque sorte ceux de l'architecture TABT. Ils comprennent :

- la comparaison de chaque item de la chaîne *input* avec les items lexicaux du système d'analyse du langage source. Moins il existe de formes « inconnues », plus grande est la probabilité de succès. Néanmoins, plus grand est le degré d'ambiguïté (défini par le nombre de sens associés à l'item du LS), plus faible est la probabilité de succès ;

- la comparaison des séquences de parties du discours avec les règles du module qui fournit les segments phrasaux (*chunker*). Plus le degré de couverture est élevé, plus grande est la probabilité de succès. Mais, à nouveau, plus grand est le degré d'ambiguïté de la couverture, moins élevée est la probabilité de réussite (voir la discussion de la TABT) ;

- l'analyse de l'importance des ressources de connaissances associées aux items lexicaux sources et utilisées pour désambiguïser ces items et construire leur représentation sémantique. Plus complètes sont ces connaissances, plus grande est la probabilité de succès d'application de la TABC pour la traduction.

Les diagnostics proposés ci-dessus varient en coûts, comme en termes de développement des procédures et de leur complexité informatique. Relativement peu coûteux sont les diagnostics basés sur la reconnaissance des formes individuelles ou des

schémas de l'entrée (contrôle des items dans un lexique ou des ressources *on-line,* contrôle de la longueur des sous-segments, contrôle des séquences de schémas de formes). Plus coûteux sont les diagnostics basés sur l'assignation de catégories aux formes. Mais, et ce à leur avantage, ces derniers sont généralement liés aux premières étapes du processus nécessaires avec la plupart des architectures, si pas toutes.

La question de comment entraîner le répartisseur va de pair avec une des questions clés de la TA moderne : comment évaluer un système de TA ?

Martin Kay a soutenu que la TA devrait prendre quelques leçons de l'ensemble du monde « statistico-connexionniste », non seulement au niveau le plus bas du système tel que le choix d'un mot, mais aussi au niveau le plus élevé. Ce qu'il entendait par cette remarque gnomique n'est pas vraiment clair, mais nous suggérons qu'un programme destiné à entraîner le répartisseur soit mis au point, programme par lequel le répartisseur serait testé sur des ensembles de textes, en assignant au hasard des parties du texte aux modules et en voyant quel type d'assignement produit les meilleurs résultats. Une variante serait qu'un humain marque intuitivement le type de texte (par le type de module qu'il juge préférable – ce qui revient à classer les textes en fonction de leur difficulté) et, à nouveau, compare ces résultats aux résultats obtenus par le système, avec les sens des traductions correctes stockées préalablement.

Administrer un tel système d'assignation de scores serait une lourde tâche. Nous envisageons ici le recours à un système du type de celui proposé récemment par Thompson (1992). Une telle méthode serait également utilisée pour évaluer la sortie du système proposé ici.

Références

BEN ARI, D., RIMON, M. et D. BERRY (1988) : « Translational Ambiguity Rephrased », *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation,* TMI-88, Pittsburgh.

BROWN, P., COCKE, J., DELLA PIETRA, S., DELLA PIETRA, V., JELINEK, F., MERCER, R. L. et P. S. ROOSSIN (1990) : « A Statistical Approach to Language Translation », *Computational Linguistics,* vol. 16, pp. 79-85.

FURUSE, O. et H. IIDA (1992) : « An Example-based Method for Transfer-driven Machine Translation », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* TMI-92, Montréal, CCRIT, pp. 139-150.

GRISHMAN, R. et M. KOSAKA (1992) : « Combining Rationalist and Empiricist Approaches to Machine Translation », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* TMI-92, Montréal, CCRIT, pp. 263-274.

ISABELLE, P. et L. BOURBEAU (1985) : « TAUM-AVIATION: Its Technical Features and Some Experimental Results », *Computational Linguistics,* vol. 11, pp. 18-27.

JONES, D. (1992) : « Non-hybrid Example-based Machine Translation Architectures », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* TMI-92, Montréal, CCRIT, pp. 163-171.

LEHMANN, H. et N. OTT (1992) : « Translation Relations and the Combination of Analytical and Statistical Methods in Machine Translation », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-92, Montréal, CCRIT, pp. 237-248.

MARUYAMA, H. et H. WATANABE (1992) : « Tree Cover Search Algorithm for Example-based Translation », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-92, Montréal, CCRIT, pp. 173-184.

MCLEAN, I. J. (1992) : « Example-based Machine Translation Using Connectionist Matching », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-92, Montréal, CCRIT, pp. 35-43.

NIRENBURG, S., CARBONELL, J., TOMITA, M. et K. GOODMAN (1992) : *Machine Translation: A Knowledge-based Approach*, San Mateo, Californie, Morgan Kaufmann.

SU, K.-Y. et J.-S. CHANG (1992) : « Why Corpus-based Statistics-oriented Machine Translation », *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-92, Montréal, CCRIT, pp. 249-262.

THOMPSON, H. (1992) : « Presentation at the AMTA Meeting », San Diego, California, November.

WILKS, Y., FASS, D., GUO, C.-M., MCDONALD, J., PLATE, T. et B. SLATOR (1990) : « Providing Machine Tractable Dictionary Tools », *Machine Translation*, vol. 5, n° 2, pp. 99-151.

15

Extraction d'un vocabulaire bilingue : outils et méthodes

Deryle LONSDALE

Center for Machine Translation, Université Carnegie Mellon, Pittsburgh, États-Unis

• *Abstract* •

This paper discusses our approach to the task of establishing a bilingual vocabulary for subsequent use in an automatic MT system. We outline methods for identifying, constraining, and aligning English source and French target vocabularies from pre-existing translation archives: source text analysis, single-word and phrasal context indices, identification of nominal compounds and variants, integration of client terminology resources, automatic alignment of source and target texts, extraction of translated source terms, vocabulary revision, and conversion for MT lexicon purposes. We also describe the tools designed and developed to help us in this effort: indexers, monolingual and bilingual context browsers, parsers, programs for describing nominal compound constituency, and a translation editor tool.

Introduction

Parmi les différentes approches de la traduction automatique ou automatisée figure le modèle dit *à base de connaissances.* Certains essais préliminaires, comme par exemple le projet KBMT-89 (Goodman et Nirenburg 1991) et le système apparenté KANT (Nyberg et Mitamura 1992), utilisent, pour atteindre une traduction de haute qualité et un traitement automatique, une analyse sémantique relativement poussée. Cette analyse se base sur une représentation explicite des concepts et des relations qui les lient.

Depuis longtemps déjà, on sait qu'il existe une différence appréciable entre la création d'un prototype expérimental d'une part et l'élaboration d'un système de TA ou de TAO utile et convivial, d'autre part. Dans le cas des systèmes de traduction

basés sur les connaissances, la commercialisation exige une saisie massive de données, qui requiert une collaboration étroite entre client et fournisseur, facilitée par une technologie avancée. D'après Galinski (1988), le problème essentiel des chercheurs dans ce domaine est, en effet, d'obtenir des spécialistes techniques des données complètes et fiables pour chaque concept du domaine, de garder ces données à jour et de les restructurer automatiquement.

Ces derniers temps, on discute beaucoup de l'utilité des ressources textuelles en ligne et des moyens de les exploiter pour l'acquisition des connaissances. Même si tout ne peut se faire automatiquement, il faut réduire le temps de participation des experts humains, tout en facilitant autant que possible leur travail. Ainsi, le coût de développement sera moindre et l'expert (en traduction, terminologie, ou un domaine technique) ne sera pas trop dépaysé par l'apport d'informatique.

Nous montrons ici comment nous avons élaboré les ressources lexicales requises pour réaliser un système capable de traduire en français une quantité importante de documents techniques.

Nous ne traiterons pas ici de l'acquisition des autres formes de connaissances requises par le système, par exemple :

- l'agencement du modèle hiérarchique des concepts du domaine ;
- les règles grammaticales de structure syntaxique des langues source et cible ;
- les règles de restriction et de composition de l'interpréteur sémantique ;
- les règles de mise en correspondance qui assurent la réalisation de concepts neutres en langue cible.

Toutefois, les principes qui ont régi l'acquisition des données lexicales et les outils utilisés nous ont également servi dans ces autres tâches.

Identification du vocabulaire source

Dans le cadre du projet KANT, nous développons un système de traduction capable de traiter des textes rédigés en anglais contrôlé. Il nous a donc fallu fixer le vocabulaire anglais source du domaine traité : la documentation technique pour les équipements lourds. Ainsi, nous avons commencé par chercher à définir avec l'aide du client l'ampleur de vocabulaire et la complexité syntaxique requises pour permettre aux auteurs de s'exprimer aussi clairement et simplement que possible.

D'abord, nous avons dû évaluer les types de documents publiés, leur degré d'adéquation aux projets et leurs traits principaux. Heureusement, le client possédait déjà des renseignements utiles concernant ces questions ; nous avons ainsi pu caractériser le domaine source, sans recourir à une analyse détaillée par des techniques d'extraction d'informations. Le domaine particulier que nous avons décidé de traiter comprend donc les manuels de fonctionnement et d'entretien, les listes de pièces détachées et les revues d'entretien, et ce pour plusieurs produits différents.

Nous avons reçu du client un corpus de plusieurs centaines de fichiers, comprenant quelque 53 megaoctets de textes préalablement publiés, qui représentent le

domaine choisi. Ces textes contenaient une quantité importante de tableaux, dessins, figures, énumérations et diagrammes, avec plusieurs codes de mise en page et de typographie, provenant des différents systèmes utilisés. Nous avons donc développé un ensemble de programmes et outils pour standardiser le contenu linguistique des textes, ce qui nous a notamment permis de normaliser l'espacement entre mots, phrases, paragraphes et mesures, ainsi que la ponctuation, les caractères spéciaux (à 8 bits, etc.), l'emploi des lettres majuscules et les codes les plus importants de composition. Ces outils consistent d'une part en des sous-programmes UNIX (AWK, LEX, SED, etc.), et d'autre part, en des programmes codés en LISP et C.

Nous avons ainsi obtenu un corpus de fichiers de textes anglais relativement restreint : les 7 millions de mots (environ) ne contiennent que 12 000 mots uniques, dont une quantité importante de numéros de pièces, d'acronymes, d'initiales et de noms propres.

Nous avons ensuite développé un indexeur codé en C qui enregistre la position de tous les mots simples qui figurent dans le corpus. Cet index sert, comme nous le verrons bientôt, de ressource pour les outils de contextes unilingues et bilingues.

Lorsqu'il est associé à un étiqueteur de codes grammaticaux, cet indexeur est aussi capable d'identifier et d'indexer des structures lexicales, ce qui nous a permis de trouver et de stocker les composés nominaux[1], les groupes verbe + préposition, et d'autres collocations. Pour ce faire, nous nous sommes servis du corpus Brown, mais nous avons dû y apporter certaines modifications. Par exemple, plusieurs mots techniques ne figurent pas dans ce corpus de référence : *screed, zener,* etc. En revanche, plusieurs catégorisations générales ne sont pas appropriées dans ces domaines techniques, où, par exemple, *keep* et *will* ne sont plus des nominaux. Moyennant de telles modifications, ce processus a permis de récupérer plus de 110 000 occurrences de composés nominaux distincts (y compris les flexions morphologiques).

Nous avons ensuite procédé à l'élimination des formes jugées inutiles, comme les termes qui présentent des fautes de frappe ou une orthographe non standard et certaines abréviations :

- *exhaust gasses = exhaust gases*
- *fuel sulphur content = fuel sulfur content*
- *9 tooth dog clutch = nine-tooth dog clutch*
- *digital lcd display = digital liquid crystal display*

La normalisation de l'espacement, surtout avec l'emploi du trait vertical et du trait d'union, a permis de mettre en relation plusieurs centaines de termes :

- *air to air aftercooler = air-to-air aftercooler*
- *blowby/airflow instrument = blowby/air flow instrument*
- *one way clutch = one-way clutch*
- *infra-red light = infrared light*
- *air/fuel mixture = air-fuel mixture*

1. Par *composé nominal,* nous entendons les collocations de nominaux qui forment la lexicalisation d'un concept du domaine, ce qui inclut les syntagmes lexicaux et souvent les syntagmes libres (voir Bédard 1986).

L'identification de variantes dérivationnelles ou flexionnelles a également réduit sensiblement le nombre de termes :

- *vibratory motor = vibrator motor = vibration motor*
- *air condition system = air conditioner system = air conditioning system*
- *operator platform = operators' platform = operators platform = operator's platform*

Plus importantes sont les formes obtenues par ellipse, surtout suite à la présence dans le contexte du référent en question : la forme pleine se présente au début, suivie d'une ou de plusieurs variantes du même terme, que l'anglais contrôlé n'admet pas :

- *alkyd paint = alkyd type paint*
- *rops/fops structure = rops/fops protective structure*
- *aec switch system problem = aec pressure switch system problem*

Les composés synonymes se forment quelquefois avec les mêmes mots, mais dans un ordre différent :

- *stabilizer/dozer control main valve = stabilizer/dozer main control valve*
- *idle timer shutdown feature = idle shutdown timer feature*

Bien sûr, l'indexeur a identifié au début plusieurs formes erronées, en raison de leur ambiguïté. Par exemple, l'ambiguïté anglaise nom/verbe menait à l'extraction de termes incorrects (mis ici en gras) dans les contextes suivants :

- *Techniques develop as the* ***operator gains knowledge*** *of the truck.*
- *Apply soda solution to the battery until the cleaning action of the* ***soda stops****.*

Nous avons essayé de faire ce tri entre termes utiles/inutiles le plus automatiquement possible, avec des algorithmes comme **l'edit-distance**. Pourtant, nous avons souvent dû faire appel au corpus ou même à un expert du domaine afin d'évaluer un terme suspect. Par exemple, l'outil KWIC (décrit ci-après) montre que, contrairement à ce que nous avions soupçonné, *bendix drive* est une pièce, tandis que *executive drive* n'est qu'une rue dans une adresse. Par contre, un expert compétent unilingue a dû être consulté dans d'autres cas, surtout ceux qui concernent les synonymes et le sens des acronymes. Suite à la normalisation d'usage pour toutes ces formes, la neutralisation des formes variantes, et le rejet des formes fausses, nous avons aujourd'hui un vocabulaire de base de quelque 60 000 termes.

La tâche suivante consistait à déceler la structure des composés nominaux. Nous avons utilisé une approche plutôt conservatrice, estimant que ces composés sont formés d'une manière strictement binaire et compositionnelle (voir Levi 1978). Ainsi, à chaque niveau de composition, il n'est possible de combiner que deux sous-unités. Une sous-unité n'est admise que si elle se manifeste indépendamment ailleurs dans le corpus.

Parfois, une seule possibilité se présente, au dépens de toutes les autres :

(*auxiliary* ((*fuel filter*) (*housing assembly*)))

Pour calculer une préférence en cas de décomposition ambiguë, nous avons employé un score basé sur la fréquence d'occurrence des sous-unités, comme dans les deux exemples suivants :

```
((bypass valve displacement)
  ((((bypass valve) displacement) 98%)
    ((bypass (valve displacement)) 2%)))

((spray system water tank level)
  (((((spray system) (water tank)) level) 80%)
    ((((spray system) water) (tank level)) 12%)
    (((spray (system water)) (tank level)) 4%)
    ((spray ((system water) (tank level))) 4%)))
```

qui se traduisent respectivement par *capacité de la soupape de dérivation* et *niveau du réservoir d'eau du circuit de pulvérisation.*

Parfois, le système ne produit aucune décomposition ; ceci arrive souvent avec des formes idiomatiques et quand il y a ellipse de sous-unité. Warren (1978) note que c'est précisément ce genre de composé qui ne suit pas l'hypothèse de composition binaire.

Le corpus anglais nous a aussi servi de référence pour déterminer la stylistique et les conventions métalinguistiques, ainsi que le registre et le niveau de vocabulaire. Ces constatations se sont faites manuellement.

En somme, le corpus source est apparu comme une ressource capitale dans notre effort d'identifier et de classer le vocabulaire d'origine et celui de trouver les unités lexicales intéressantes.

Identification du vocabulaire cible

Nous avons dû ensuite identifier le vocabulaire français. Nous avons suivi à peu près la même démarche que pour le vocabulaire anglais, en nous servant des mêmes outils. Plus modeste, le corpus français ne représente que 10 % du corpus anglais.

L'extraction des codes de mise en page s'est révélé un peu plus complexe, vu l'abondance de codes différents pour les lettres accentuées et le besoin de les normaliser. Une simple extension à nos outils d'analyse source a néanmoins suffi.

Nous avons ensuite compilé les index pour les mots et syntagmes cible, après avoir préalablement amélioré l'indexeur pour tenir compte de la syntaxe de ces syntagmes. En effet, nous ne disposions pas au début de toutes les données nécessaires pour l'étiquetage du corpus français, mais nos outils lexicographiques nous ont aidés dans cette tâche. L'indexeur a trouvé quelque 19 000 mots simples et 160 000 composés nominaux dans le corpus français, qui compte quelque 1 750 000 mots.

Le vocabulaire français a pu, lui aussi, être normalisé. Puisque les résultats de cet effort sont analogues à ceux de l'anglais, nous ne donnerons pas d'exemples.

Notre outil KWIC (*KeyWord in Context* : mot-clé et contexte, figure 1) est conçu pour afficher le contexte de n'importe quel mot ou syntagme figurant dans les index des corpus source et cible. Il est suffisamment général pour traiter l'anglais, ainsi que les autres langues envisageables dans un avenir proche. L'outil permet de voir les occurrences séquentiellement ou dans des contextes triés (à gauche ou à droite), avec longueur de contexte variable. Puisque cet outil fonctionne sur une gamme de plates-formes logiciel/matériel, nous avons dû restreindre sa fonctionnalité d'affichage et d'interaction.

```
Term to show? antivandalisme
iller A bien verrouiller le bouchon      antivandalisme  .\vErifications extErieures\contr0l
iller A bien verrouiller le bouchon      antivandalisme  .\vErifications extErieures\contr0l
ous et des couvercles de protection      antivandalisme  .\7 .\raccrocher la ceinture au sup
ssent plus facilement de protection      antivandalisme  .\les rayures fines et les petites
5 .\verrouiller tous les couvercles      antivandalisme  et tous les compartiments ( coffre
protections et tous les couvercles       antivandalisme  .\conseils pour le transport\expEdi
e les protections et les couvercles      antivandalisme  .\conseils pour le transport\expEdi
e les protections et les couvercles      antivandalisme  .\rEglages\abaissement de la flEche
e les protections et les couvercles      antivandalisme  .\conseils pour le transport\expEdi
ler tous les bouchons et couvercles      antivandalisme  ( coffre A outils , capot du moteur
ndie .\3 .\monter tous les bouchons      antivandalisme  et couvercles .\conseils pour le tr
la vis de fixation des protections       antivandalisme  . extraire les protections .\7 .\mo
uvercles et verrous des protections      antivandalisme  \6 .\pour descendre du vEhicule ,
e ; mettre en place les protections      antivandalisme  .\7 .\obturer l'Echappement afin d'
e ; mettre en place les protections      antivandalisme  .\7 .\obturer l'Echappement afin d'
nt ) .\6 .\dEtacher les protections       antivandalisme  A l'avant du radiateur , sur le c0t
tterie .\4 .\monter les protections       antivandalisme  et les couvercles .\transport de la
achine .\7 .\monter les protections       antivandalisme  sur le pare-brise et la portiEre ;
ctions .\7 .\monter les protections       antivandalisme  sur le pare-brise et les verrouille
tre en place toutes les protections      antivandalisme  . fermer et verrouiller toutes les
.\5 .\monter toutes les protections      antivandalisme  .\6 .\pour descendre du vEhicule ,

Term to show? refroidisseur d'huile hydraulique
transformation 8e3162 de   refroidisseur d'huile hydraulique   pour remplacer le faiscea
8\canalisations 9c1105 de  refroidisseur d'huile hydraulique   \8/1d/7x7958 A la place de
teur de climatiseur et de  refroidisseur d'huile hydraulique   n'interviennent pas dans
sceaux de radiateur et de  refroidisseur d'huile hydraulique   .\pour monter cette grill
circuit hydraulique .\le   refroidisseur d'huile hydraulique   9c2293 comprend un faisce
le de transformation pour  refroidisseur d'huile hydraulique   /eq/ pour tracteurs d6h (
le de transformation pour  refroidisseur d'huile hydraulique   \1375 , 1379\tracteurs d6h
le de transformation pour  refroidisseur d'huile hydraulique   \tracteurs d6h ( 1fj , 1kd
ton . refermer le capot .\ refroidisseur d'huile hydraulique   \nettoyage\situE sur le c0
/dl/ 65\nettoyage /dl/ 65\ refroidisseur d'huile hydraulique   /dl/ 67\nettoyage /dl/ 67
ent /ld/ vErifier /dl/ 67\ refroidisseur d'huile hydraulique   /ld/ nettoyer /dl/ 67\rEs

Term to show? □
```

FIGURE 1 : Affichage de l'outil KWIC.

Appariement bilingue des vocabulaires

Puisque la collection largement automatique des vocabulaires était requise pour d'autres raisons, nous avons choisi d'associer autant que possible les unités lexicales des deux langues suivant un processus d'appariement direct.

Pour commencer, nous avons réuni toutes les ressources lexicales bilingues du client. Celui-ci avait à sa disposition une banque de données bilingues utilisée pour remplir les commandes de pièces détachées. Comme c'est souvent le cas avec de telles ressources, le format des entrées ne convenait pas complètement ; un remaniement avec des programmes LISP et une révision manuelle se sont imposés. Le client a aussi fourni un modeste lexique bilingue de termes simples, avec annotations, compilé par plusieurs traducteurs. Ce lexique, lui aussi, a nécessité une réorganisation presque totale. L'emploi de plusieurs dictionnaires bilingues en ligne nous a fourni quelques traductions pour chaque terme source général (simple et composé). Nous avons rejeté automatiquement les traductions qui ne figurent jamais dans le corpus cible. L'intégration de ces trois ressources lexicales bilingues nous a servi de point de départ. Toutefois, de cette manière, nous n'avons pu récupérer que quelque 6 500 termes, soit 11 % des termes source.

Nous avons ensuite entrepris une autre étape de collection d'équivalents source/cible, basée sur les deux corpus respectifs. Nous avons développé un programme d'alignement des corpus qui nous a permis d'accéder au contexte de tout terme source et de trouver le contexte correspondant du texte traduit dans un fichier de langue cible. À l'aide des index de termes source, nous avons trouvé chaque occurrence du terme et scanné les fichiers cible pour retrouver les contextes correspondants en français. Nous avons conçu et développé un outil sous le système d'interfaces X-Windows pour afficher les résultats de l'alignement. Ce dernier soutient aussi l'interaction avec un utilisateur, ce qui permet à ce dernier de sauvegarder dans une base de données les correspondances source/cible intéressantes.

L'alignement n'a guère posé de problèmes, contrairement à notre attente. Malheureusement, en raison de divergences importantes et fréquentes entre les textes anglais et français, une approche basée sur les algorithmes statistiques à programmation dynamique (comme Gale et Church 1991) n'a pas produit d'excellents résultats. Par contre, nous avons développé notre propre approche qui suit l'esprit de Simard *et al.* (1992), comme nous l'avons remarqué plus tard.

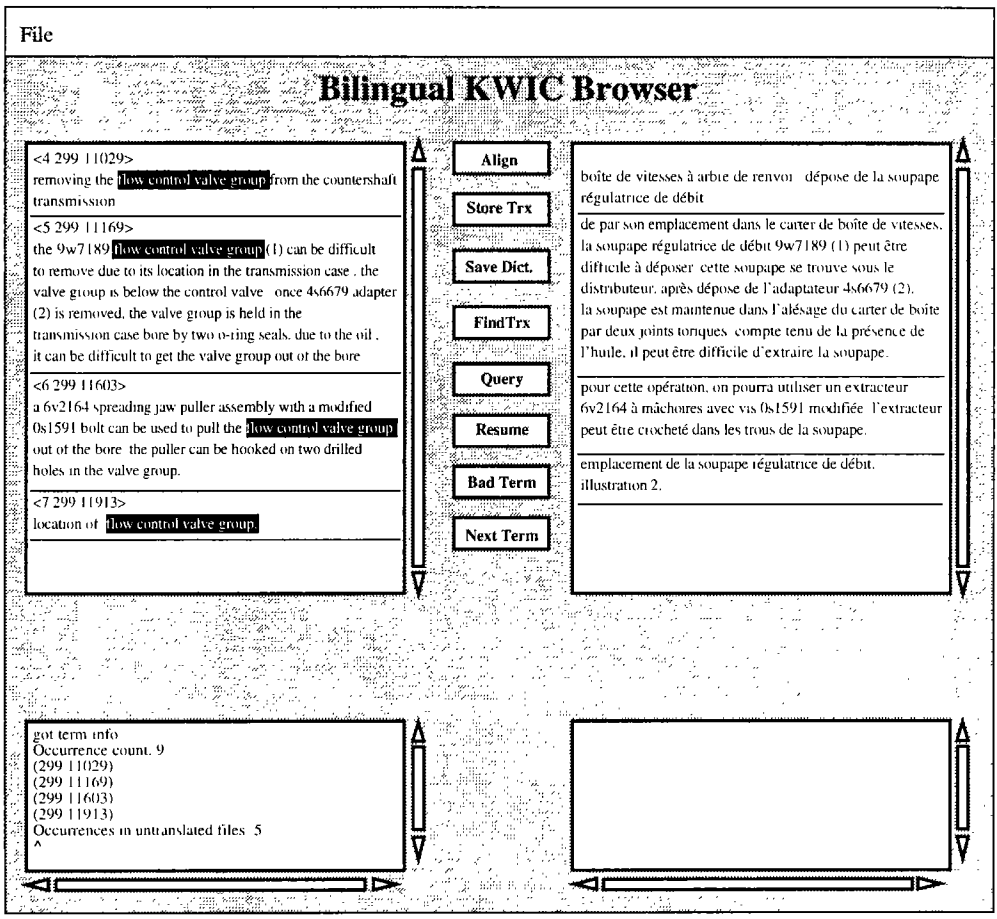


FIGURE 2 : Affichage de l'outil BiKWIC.

Nous avons tiré profit de certains éléments du texte qui restent invariables dans la traduction. Par exemple, le texte est particulièrement riche en diagrammes étiquetés qui ne changent jamais de langue en langue et dont les renvois restent invariables dans le texte. Il contient aussi une quantité importante de mesures, chiffres et numéros de pièces. Chaque phrase est donc classifiée d'après les points saillants les plus distinctifs de son contenu : mesures, renvois numériques, numérotation, et catégorisation lexicale générale (ce dernier critère étant le plus lent à tester et donc le moins usité). Une comparaison des codes de classification de phrases permet de faire défiler de façon parallèle les deux fichiers, combinant les phrases source ou cible au besoin.

Nous avons trouvé des divergences intéressantes à la suite de ce processus. Même si la terminologie est assez difficile en raison de son caractère technique, nous avons trouvé que les personnes qui ont une bonne connaissance générale de la langue peuvent au moins isoler et identifier ces unités terminologiques. Ainsi, nous avons employé à temps partiel plusieurs étudiants bilingues pour récolter ces termes dans les alignements.

Se servant de l'outil BiKWIC (figure 2), ils prennent une liste de termes source à traduire et examinent les contextes alignés dans les corpus source et cible. Lorsqu'ils trouvent dans ces contextes une correspondance traductionnelle, ils enregistrent la traduction avec une opération de souris dans une base de données terminologiques associée à l'outil, qui sera revue plus tard par un traducteur/terminologue expert dans le domaine. De cette façon, nous avons pu extraire les traductions de quelque 8 000 composés nominaux et de 4 000 mots simples. Un traitement plus sophistiqué d'alignement intra-phrastique nous permettrait d'extraire de façon automatique ces unités lexicales et leurs traductions ; nous espérons en faire mieux la preuve.

Toute cette démarche nous a donné un premier ensemble substantiel de vocabulaire, tiré des deux corpus reçus du client. Enfin, cette liste a été revue et les termes manquants fournis manuellement. Pour faciliter cette tâche énorme, nous avons conçu et construit un outil interactif pour l'enregistrement de ces termes. Un traducteur expert dans le domaine s'est servi de l'outil pour faire ces deux tâches à la fois.

Cet outil de rédaction de vocabulaire affiche les termes anglais, un à un, avec tous les termes apparentés (c'est-à-dire ceux qui ont au moins un mot en commun). Le cas échéant, les traductions du terme vedette sont aussi affichées. Pour chaque terme à traduire, le système essaie de proposer une traduction brouillon, en comparant celles des termes apparentés pour en tirer les généralités nécessaires à la construction par inférence d'une traduction. Les termes traduits les plus utiles sont ceux qui incluent le terme vedette comme sous-terme : un découpage de la traduction du terme entier s'impose dans ces cas.

Recodage en forme fonctionnelle

Une fois la correspondance lexicale établie entre les concepts neutres du domaine et les unités lexicales de la langue cible, il faut recoder celles-ci dans une forme plus fonctionnelle. Le générateur requiert une description de la structure syntaxique de ces unités ; sinon, les flexions et autres modifications risquent de se faire incorrectement.

On peut considérer, à titre d'illustration, le problème de la représentation lexicale des composés nominaux français. Le sous-système exécuté pour générer le texte français de sortie suit un ensemble de règles grammaticales qui décrivent la morphologie et la syntaxe de la langue cible. Il opère sur des données que l'on appelle des *structures fonctionnelles grammaticales* (des SFG), les traitant selon un algorithme récursif et descendant. Chaque unité lexicale doit donc être recodée en SFG pour pouvoir être intégrée avec les autres constituants de la phrase. Ces SFG, qui peuvent d'ailleurs devenir plutôt complexes, peuvent, elles aussi, être créées automatiquement et révisées au besoin.

La formation des SFG se fait à l'aide de notre parseur de texte français. Ce parseur comprend l'analyseur-compilateur généralisé gauche-à-droite qui convertit un ensemble de règles syntagmatiques de la grammaire française en une table d'analyse. Ces règles suivent de très près le format de celles du générateur. (Ces deux grammaires du français sont proches mais non pas identiques, ne supportant donc pas un traitement pleinement réversible analyse/génération.) Le parseur prend ainsi la forme lexicale du terme français pour le convertir en SFG.

Pendant cette conversion texte-SFG, nous avons rencontré les problèmes classiques : attachement des syntagmes prépositionnels, adjectifs, etc., ce qui cause souvent la multiplication des représentations candidates. Par exemple, si le terme *front main control valve bank* se traduit *bloc de distributeur principal avant,* il faut attacher correctement le mot *avant* pour représenter la structure correcte du terme français. Parmi les dix compositions admises par le corpus, celle qui est jugée la plus acceptable d'après nos métriques se forme ainsi :

(*front* ((*main* (*control valve*)) *bank*))

Admettant, d'après Bauer (1978), une généralisation parallèle en français, on obtient la SFG suivante :

```
( (CAT noun) (ROOT bloc) (POSTMOD -)
 (AGR
    ( (NUMBER sg) (GENDER m) ) )
 (MODIFIER
    ( (CAT adj) (ROOT avant) ) )
(PP
  ( (P-OBJ
      ( (ROOT distributeur) (CAT noun)
        (AGR
           ( (NUMBER sg) (GENDER m) ) )
        (MODIFIER
           ( (CAT adj) (ROOT principal) ) ) ) )
  (PREP
     ( (ROOT de) (CAT prep) ) ) ) ) )
```

Nous avons dû intervenir pour de tels attachements ambigus ; heureusement, nous avons pu nous aider du corpus des décompositions pour l'anglais.

Ces unités lexicales et leurs formes SFG sont utilisées dans le système par le mappeur, qui assure la réalisation des concepts primitifs interlangues en structures linguistiques de la langue cible. Nous ne pouvons pas entamer ici une discussion de cette étape ; il suffit de noter que la collection de toutes ces classes de données requises par le mappeur et le générateur constitue une partie essentielle du système.

Améliorations futures

Durant les travaux du projet discuté ici, nous avons dû mettre en balance les techniques automatiques et parfois expérimentales et les approches reconnues mais moins informatisées, ce qui nous a parfois forcés à remettre à plus tard un traitement de pointe prometteur mais qui n'a pas fait ses preuves. Dans ce chapitre, nous mentionnons certaines des améliorations possibles.

Notre liste étiquetée de mots anglais ne reflète pas complètement la nature technique du corpus source. Pour cette raison, les structures lexicales contiennent parfois des erreurs qui nécessitent une révision manuelle. Un meilleur ensemble d'étiquettes,

ou une heuristique d'induction de catégories grammaticales pourrait nous aider à ne retirer que les unités lexicales intéressantes.

Nous n'avons pas utilisé de techniques de troncature, comme celles traitées dans la littérature d'extraction (par exemple Savoy 1991), pour neutraliser les flexions des formes morphologiques. Pour des raisons pragmatiques, nous avons développé des algorithmes et programmes *ad hoc* ; ils ont été utiles dans notre cas, mais ils ne se généralisent pas particulièrement bien. Nous comptons intégrer une approche plus satisfaisante.

Nous allons aussi améliorer les heuristiques d'identification des structures lexicales intéressantes. Par exemple, nos sous-programmes, écrits actuellement en C, peuvent être extraits avec l'indexeur et recodés en une langue telle que LEX qui admet plus facilement une spécification de structures syntagmatiques. Nos outils seront alors plus facilement extensibles.

Notre technique d'alignement bilingue, bien que satisfaisante, est tributaire du caractère spécial d'une sous-classe des textes du client. Si nous comptons, à l'avenir, traiter des textes sans mesures, ni renvois, etc., il nous faudra y intégrer les techniques récemment décrites, par exemple, dans Church (1993).

L'extraction d'équivalents ici n'est pas totalement automatique : nous manquons en effet de ressources lexicales primaires pour ce projet et les techniques d'appariement lexical sont toujours sous investigation. Debili et Sammouda (1992) présentent une solution possible.

Nos premiers essais en description compositionnelle ont fourni une quantité sensible de données, mais nous n'avons pas encore trouvé la façon de tirer pleinement parti de cette ressource dans la spécification bilingue des unités lexicales. Néanmoins, cette démarche semble riche en possibilités.

Au cours des efforts de développement du vocabulaire, nous avons souvent dû consulter un expert dans le domaine approprié. Même des questions de synonymie, de polysémie, et d'usage non standard ont parfois nécessité ces interventions lentes et difficiles, mais aussi coûteuses. Nos outils KWIC et BiKWIC nous ont souvent aidés à résoudre certains de ces problèmes ; toutefois, il est sans doute possible de relever davantage de renseignements directement des corpus par des moyens plus automatiques.

Finalement, il a souvent été difficile de quantifier au préalable nos efforts en chiffres exacts : temps de traduction en fonction de nombre de termes, grandeur des corpus et résultats de l'analyse, ampleur du vocabulaire final et des formes à rejeter, ressources informatiques nécessaires, etc. Le développement de meilleures techniques d'estimation et de quantification devrait à l'avenir réduire le coût et les risques associés au développement de vastes bases de connaissances.

Conclusion

Nous avons décrit ici un ensemble d'outils et de méthodes d'acquisition de la connaissance lexicale, mis au point pendant le développement d'un système de traduction à

pivot interlingue. Sa mise en application n'aurait guère été possible sans ces outils, vu le goulot d'étranglement dû au volume énorme de connaissances nécessaires. Nos méthodes se basent autant que possible sur une analyse automatique des corpus source et cible, même si nous n'adoptons pas d'office les nouvelles techniques de pointe encore en développement.

Il apparaît clairement qu'en attendant la preuve de certaines technologies textuelles, il est déjà possible d'intégrer l'informatique et l'expertise humaine pour extraire le contenu de tels corpus. Même la vaste tâche d'élaboration d'une base de connaissance compréhensible pour tout un domaine technique n'échappe pas à une telle collaboration.

Cette dernière exige au moins une panoplie d'outils qui assurent le remaniement des fichiers de textes, l'analyse grammaticale et statistique des mots et phrases, l'affichage de termes dans leurs contextes unilingues et bilingues et l'interaction efficace avec les experts humains. Notre contribution montre que de telles entreprises sont faisables et que ces moyens peuvent augmenter appréciablement les efforts en lexicologie, terminologie et traductique.

Remerciements

Je tiens à remercier tous mes collègues du projet KANT, surtout John Leavitt et Eric Nyberg du CMT, et Claude Doré de Traductions Taurus.

Références

BAUER, L. (1978) : *The Grammar of Nominal Compounding with Special Reference to Danish, English, and French,* Odense University Press.

BÉDARD, C. (1986) : *La traduction technique : principes et pratique,* Montréal, Linguatech.

CHURCH, K. W. (1993) : « Char_align: A Program for Aligning Parallel Texts at the Character Level », *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics.*

DEBILI, F. et E. SAMMOUDA (1992) : « Appariement des phrases de textes bilingues », *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92,* Nantes.

GALE, W. et K. CHURCH (1991) : « A Program for Aligning Sentences in Bilingual Corpora », *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* Berkeley, June 18th-21st.

GALINSKI, C. (1988) : « Advanced Terminology Banks Supporting Knowledge-Based MT », Maxwell, D., Schubert, K., et Witkam, A. P. M. (dir), *New Directions in Machine Translation,* Foris Publishers.

GOODMAN, K. et S. NIRENBURG (dir) (1991) : *A Case Study in Knowledge-based Machine Translation,* Morgan Kaufmann.

LEVI, J. N. (1978) : *The Syntax and Semantics of Complex Nominals*, New York, Academic Press.

NYBERG, E. et T. MITAMURA (1992) : « The KANT System: Fast, Accurate, High-quality Translation in Practical Domains », *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes.

SAVOY, J. (1991) : « Stemming of French Words », Département d'informatique et de recherche opérationnelle 793, Université de Montréal.

SIMARD, M., FOSTER, G. F. et P. ISABELLE (1992) : « Using Cognates to Align Sentences in Bilingual Corpora », *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, TMI-92, Montréal, CCRIT.

WARREN, B. (1978) : *Semantic Patterns of Noun-Noun Compounds*, Acta Universitatis Gothoburgensis.

16

La traduction de prépositions temporelles

Siety MEIJER

CL/MT Group, Université d'Essex, Wivenhoe Park, Angleterre

• *Abstract* •

The translation of prepositions poses a problem for MT, and temporal prepositions are no exception. The problems predominantly lies in those prepositions that have little semantic content (e.g. "at" and "in", but not "during" and "before"). When looking at a possible solution for this problem I will concentrate on a lexicalist approach. This approach uses the idea introduced originally by Mel'čuk to lexicalise semantically empty elements. He introduces lexical functions of which LOC_{in}^{temps} *is highly relevant. By specifying for each noun the appropriate temporal localizer we can solve most problems in the translation of temporal prepositions. An important question which this approach needs to answer is whether or not to elevate the semantically empty preposition and delete the prepositional node.*

Introduction

La traduction des prépositions pose un problème essentiel pour la TA : celui de l'ambiguïté. La plupart d'entre elles, en effet, peuvent recevoir plusieurs équivalents traductionnels et les prépositions temporelles ne font pas exception. Dans cet article, nous traiterons des problèmes posés par les constructions GP-GN, sans prendre en considération les connecteurs temporels. Ces problèmes sont bien illustrés dans les exemples suivants (anglais, hollandais et espagnols) :

1. *He woke up*

a) *at 5 o'clock*	*vijf uur*	*a las cinco*
b) *at night*	*'s nachts*	*por la noche*
c) *at the weekend*	*in het weekend*	*durante el fin de semana*

2. *We played football*			
a) *on Monday*	*(op) maandag*	*el lunes*
3. *They discussed it*			
a) *in the evening*	*'s avonds*	*por la noche*
b) *in the summer*	*in de zomer*	*en el verano*
c) *in the meeting*	*in de vergadering*	*durante la reunion*

Pour aborder cette question, différentes méthodes ont été envisagées. Nous commencerons par présenter la méthode de Brée *et al.* (1990 ; 1992) et celle qui a été adoptée par le groupe de recherche EUROTRA pour le temps et l'aspect (Meijer *et al.* 1992 ; 1993) et nous mettrons en évidence leur principal désavantage. Nous introduirons ensuite une méthode lexicale qui permet un traitement plus adéquat de ce problème de traduction. Le dernier chapitre montrera comment mettre en œuvre cette méthode.

Description de sens des prépositions temporelles

Une des solutions possibles à notre problème est de définir une description sémantique unique pour tous les sens différents des prépositions temporelles et de les mettre en correspondance avec une préposition de même sens dans la langue cible. Brée *et al.* (1990 ; 1992) appliquent cette méthode pour le domaine de la compréhension des langues naturelles. Leur recherche traite des prépositions et des connecteurs temporels et tente d'établir « une description du sens des mots fonctionnels temporels telle que ces sens puissent être inclus dans la composante sémantique de tout système de traitement des langues naturelles ».

Cette méthode utilise des *arbres de sélection*, une variante des arbres de décision conventionnels. Ces arbres, définis pour l'anglais et le hollandais, devraient rendre compte de la plupart des usages des prépositions et connecteurs temporels.

Les figures 1 et 2, qui reprennent une partie de l'arbre de sélection pour l'anglais et la partie correspondante pour le hollandais (extraites de Brée *et al.* 1990), exemplifient leur fonctionnement. Les deux arbres représentent les prépositions temporelles duratives et les prépositions locatives directes (qui n'expriment pas de relation d'ordre).

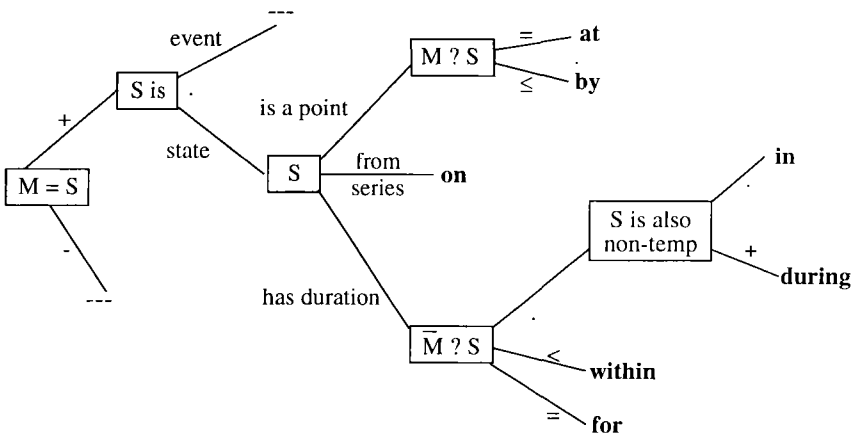


FIGURE 1 : Partie de l'arbre de sélection pour l'anglais.

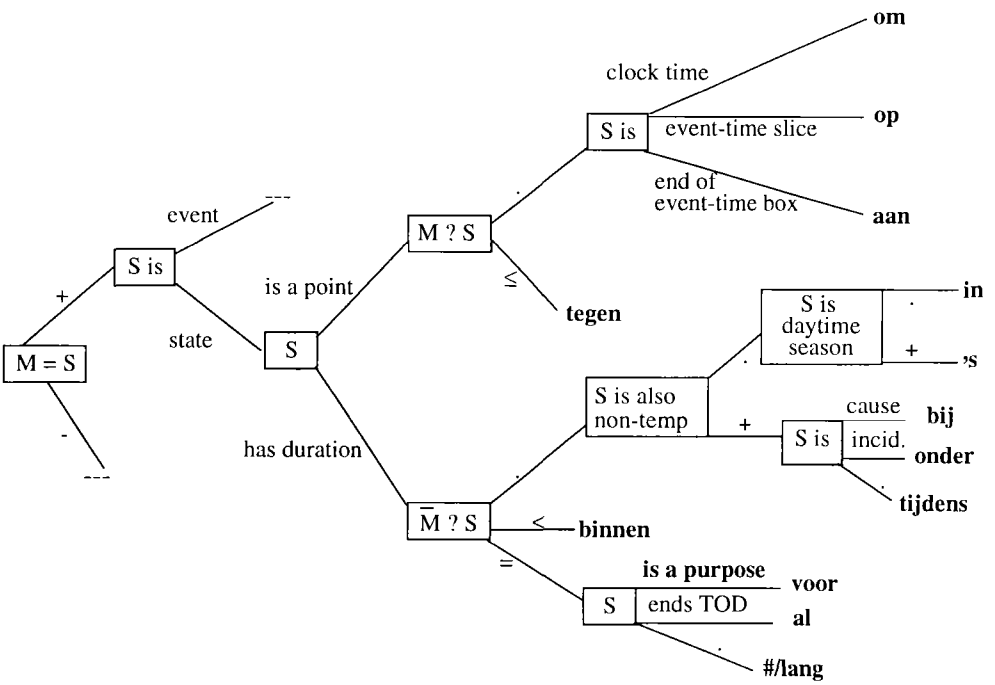


FIGURE 2 : Partie de l'arbre de sélection pour le hollandais.

Prenons, par exemple, l'arbre anglais : des décisions doivent être prises pour avancer du nœud racine au nœud feuille qui représente le mot fonctionnel temporel approprié. Pour avancer, par exemple, du nœud racine au nœud *S is,* il faut calculer si le temps du verbe (M) équivaut à celui du modificateur temporel ou sous-état (S) – l'autre solution étant que le modificateur temporel exprime une relation d'ordre, comme c'est le cas pour *after* ou *since.* La distinction suivante concerne le modificateur temporel et son caractère événementiel, ce qui permet de distinguer les connecteurs temporels (sans relation d'ordre) des prépositions temporelles locatives et duratives. L'étape suivante vérifie si le modificateur temporel réfère à un instant précis, un intervalle ou s'il fait partie d'une série (comme les jours). Pour les modificateurs qui se rattachent à un instant précis, il faut faire une distinction supplémentaire : le sous-état est-il simultané avec le verbe (par exemple *at 8 o'clock*) ou le verbe exprime-t-il quelque chose qui peut avoir lieu avant le sous-état (*by 8 o'clock*) ? Parmi les prépositions duratives, il faut distinguer celles qui sont simultanées (*for three hours*) des inclusives (*within 3 hours*). Le choix entre les deux autres modificateurs temporels *in* ou *during* dépend du caractère temporel du nom. L'arbre de sélection correspondant pour le hollandais est en grande partie similaire.

Cette recherche conduit à une description assez concise du comportement combinatoire des mots fonctionnels temporels, représentée sous la forme d'une structure organisée, l'arbre de sélection. Nous faisons cependant quelques réserves sur l'utilité de ces arbres en traduction automatique. Ces réticences se fondent sur les observations suivantes :

1. **Les arbres de sélection ne sont pas complets.** L'arbre de sélection de Brée (1990) ne traite pas des prépositions anglaises suivantes : *over, around, about, near, towards, through, throughout* et *under.* Certaines sont couvertes par Brée (1992), mais pas toutes. En outre, cette méthode ne décrit pas tous les usages possibles des prépositions. Nous reviendrons sur ce point dans le chapitre sur les désavantages des deux méthodes.
2. **Les arbres de sélection sont trop ambitieux pour la TA actuelle.** Ils ont été construits pour être incorporés dans un programme de TAL capable de comprendre le texte avant de le traduire. Or, la plupart des systèmes de traduction automatique actuels sont loin de pouvoir comprendre les textes et certaines des distinctions utilisées dans les arbres de sélection sont difficiles à appliquer, même pour des systèmes de compréhension de textes avancés.
3. **La méthode est trop formalisée.** Il existe environ 29 prépositions temporelles anglaises, parmi lesquelles 4 seulement (*at, in, on* et *for*) sont problématiques (du moins, en ce qui concerne la traduction en hollandais ou en espagnol). Une méthode aussi formalisée nous paraît donc peu utile.

Classification des noms temporels

Parmi les prépositions temporelles problématiques qui viennent d'être présentées, *for* est une préposition durative. Pour la traduire correctement, il faut prendre en considération le temps et l'aspect du verbe. Dans cet article, je ne traiterai que des trois autres prépositions : *at, in* et *on,* qui comptent parmi les prépositions temporelles les plus fréquentes. Leur sélection dépend du type du nom qui les gouverne, ce qui a motivé un groupe de chercheurs EUROTRA à proposer une méthode de rechange pour

celle de Brée, basée sur les traits sémantiques. Certains traits sémantiques fondamentaux, utilisés dans les entrées lexicales des noms, comme, par exemple, *semtype=temp* ou *semtype=loc,* permettaient déjà de résoudre le problème de la traduction de *at* dans *at school* (locatif) et de *at 3 o'clock* (temporel). Pour étendre cette méthode aux prépositions temporelles, le groupe de chercheurs a proposé l'utilisation d'un trait plus spécifique : **tempval**. Ce trait doit être présent dans toutes les entrées lexicales des noms temporels et peut avoir les valeurs suivantes :

tempval = *hour_of_day, part_of_day, weekday, name_of_month, name_of_season, name_of_year, name_of_century.*

assignées, par exemple, de la manière suivante :

hour_of_day :	*3 o'clock, midnight, noon, etc.*
part_of_day :	*morning, afternoon, evening, night*
weekday :	*Sunday, Monday, Tuesday, etc.*
name_of_month :	*January, February, etc.*
name_of_season :	*winter, spring, summer, autumn*
name_of_year :	*1492, 1964, 1991, etc.*
name_of_century :	*nineteenth century, etc.*

Pour pouvoir être utilisé dans le choix de la traduction correcte des prépositions temporelles, le trait *tempval* doit être communiqué au nom par la règle suivante, qui copie la valeur du trait *tempval* dans la préposition.

```
rule2 = {}[{tempval=T/cat=p}, {/cat=np}
                    [{/cat=n, tempval=T}, *{}]].
```

Une fois le trait *tempval* disponible au niveau de la préposition, il est possible d'écrire des règles de transfert simples qui permettent d'obtenir la traduction correcte des prépositions temporelles. La règle de transfert suivante traduit, par exemple, la préposition anglaise *at* en *a* (espagnol) dans le cas d'un GP du type *at midnight* (*at midnight => a medianoche*) :

```
rule3 = {gb_lu=at, tempval=hour_of_day} => {e_lu=a}.
```

Il faut cependant noter qu'on ne peut pas traiter toutes les prépositions temporelles avec des règles de transfert simples. Le cas de *on Monday*, par exemple, est plus complexe, puisque la préposition n'est pas traduite, mais effacée et qu'il faut insérer l'article défini *el.* Le trait *tempval* nous permet de généraliser, en écrivant une seule règle complexe qui s'appliquera à tous les jours de la semaine :

```
rule4 = ~:{cat=pp, role=mod}
            [~:{gb_lu=on},
            NP:{role=arg1, cat=np}
                    [N:{cat=n, tempval=weekday, msdefs=msabs}]]
=> NP:{role=mod} <N:{msdefs=msdefs}>.
```

Dans cette règle, le nœud GP et le gouverneur de ce nœud (la préposition) sont

effacés (indiqué avec le signe ~). Le GN est maintenu, mais devient le modificateur (rôle joué en langue source par le GP), plutôt que l'argument de la préposition. Le nom est traduit avec des règles simples, mais, ici aussi, il y a un changement de traits : la valeur *msabs* qui indique qu'il n'y a pas de déterminant est transformée en *msdefs* dans la langue cible, ce qui permettra de générer un article défini.

Désavantage de ces deux méthodes

Ces deux méthodes ne peuvent que résoudre certains problèmes de traduction : les règles sont en effet trop générales pour traiter tous les cas, comme en témoignent les exemples suivants :

Anglais	**Hollandais**
at 3 o'clock	*om 3 uur*
at that moment	*op dat moment*
at the beginning	*in het begin*
at dawn	*bij het ochtendgloren*
at night	*'s nachts*
at the weekend	*in het weekend*
at Christmas	*met Kerst*

Il est évidemment possible de développer ces méthodes, mais sans réelle garantie de succès, ce qui suggère le recours à une méthode lexicale. Cette suggestion est confirmée par la sélection de prépositions particulières dans les cas suivants : *at night* (vs *in the morning/afternoon/evening*) et *au printemps* (vs *en été/automne/hiver*).

Méthode lexicale

La méthode lexicale choisie se base sur Mel'čuk (1988). Ce dernier propose de lexicaliser les éléments qui n'ont pas de contenu sémantique et de définir ces éléments comme des fonctions lexicales sur la tête de la phrase. Par exemple, la fonction lexicale OPER appliquée au nom anglais *attempt* produit la valeur *make* ; la fonction lexicale MAGN appliquée au nom anglais *smoker* la valeur *heavy*. Pour les expressions temporelles, Mel'čuk introduit la fonction lexicale $LOC_{temp.in}$, qui a pour valeur une préposition qui dénote une localisation temporelle.

On peut se demander s'il est justifié de traiter des prépositions temporelles comme des éléments qui n'ont pas de contenu sémantique. Pour la majorité des prépositions, la réponse semble négative : les prépositions *since, after, throughout*, par exemple, contribuent à l'interprétation des GP. Il y a cependant un groupe de prépositions dont les membres ne constituent qu'une référence à un point (ou à un intervalle) sur l'axe temporel, sans rien ajouter d'autre au sens du GN temporel. Ainsi, les membres de ce groupe – les locatives directes – n'ont pas de spécification sémantique individuelle. Cette distinction entre locatives temporelles et les autres prépositions temporelles est attestée en latin où les modificateurs temporels ne sont introduits par des prépositions que s'il ne s'agit pas d'une localisation directe (exprimée par le cas ablatif).

Nous pouvons considérer comme prépositions locatives directes *at, in* et *on* en anglais, *a, en* et *por* en espagnol et *om, op, aan, in, bij* et *met* en hollandais et y appliquer la méthode lexicale. Cette dernière nous permettrait de dire que $LOC_{temp.in}$ (*night*) a la valeur *at,* tandis que $LOC_{temp.in}$ (*morning*) a la valeur *in*. Pour appliquer cette méthode, nous devons d'abord définir cette fonction lexicale particulière dans toutes les entrées des noms temporels du dictionnaire. Dans la notation de EUROTRA, cela signifie que nous devons ajouter une paire attribut/valeur à toutes les entrées lexicales qui contiennent le trait *sem=temp* :

monday = {cat=n,
ntype=subst,
person=third,
gb_lu=monday,
sem=temp,
loc_temp_in=on,
gb_isframe=arg0,
gb_rno=1}.

Avant d'appliquer la méthode lexicale, il est important d'examiner quel est le statut du GP qui contient une préposition locative directe. En latin, par exemple, ces GP sont des GN avec un marqueur relationnel – exprimé en latin par le cas ablatif (comme par exemple *the day before, yesterday,* etc.) et en anglais par une préposition. Il serait donc possible de décider d'effacer le nœud GP dans l'analyse, en marquant le GN avec le marqueur relationnel. Mais il est aussi valable d'opter pour un traitement semblable à celui utilisé pour les autres prépositions temporelles, c'est-à-dire maintenir le nœud P (et GP) dans l'analyse, mais remplacer la préposition par la valeur générale DILOC, comme dans la figure 3 (pour le GP temporel *at the moment*).

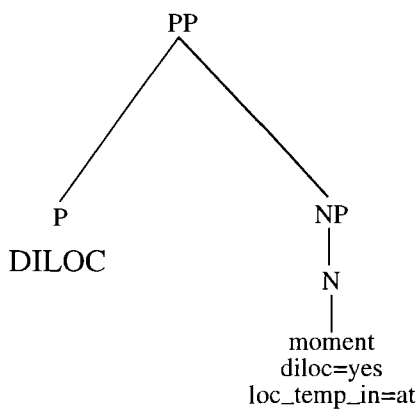


FIGURE 3 : Arbre.

Après avoir pris en considération ces deux possibilités, nous avons décidé d'adopter la deuxième et de maintenir le nœud GP. Cette décision se fonde sur des

raisons d'économie : la plupart des langues européennes discutées expriment en effet la localisation temporelle directe par une préposition.

Le processus de traduction

L'analyse

Pour l'analyse, nous avons besoin d'une règle qui identifie les prépositions temporelles locatives, remplace la préposition par la valeur lexicale générale DILOC et marque le substantif au moyen du trait *diloc=yes*. La règle suivante suffit pour l'analyse de l'anglais :

```
ta_diloc = PP:{cat=pp}
              [P:{cat=p,sf=gov,gb_lu=PREP},
              NP:{cat=np}
                    [N:{cat=n,sf=gov,loc_temp_in=PREP},
                    REST:*{}]]
=> PP < P:{gb_lu=DILOC}, NP < {cat=n, diloc=yes}, REST > >.
```

Mais cette règle ne s'applique pas à tous les cas. Comme nous l'avons déjà vu, certains modificateurs temporels qui indiquent la localisation directe sur l'axe temporel ont une préposition dans une langue, mais pas dans une autre. L'espagnol, par exemple, utilise un article défini pour traduire *on Monday* (*el lunes*). Ce problème sera discuté dans la prochaine section. Nous montrerons d'abord comment identifier et traiter les locatives directes sans préposition dans l'analyse.

```
ta_diloc2 = GOV:{cat~=pp,}
              [NP:{cat=np,sf=mod}
                    [N:{cat=n,sf=gov,loc_temp_in=nil}
                    REST:*{}]
              B:*{}]
=> GOV < NP < {cat=n, diloc=yes}, REST > B >.
```

Cette règle stipule que si un nom a le trait *loc_temp_in=nil,* si son gouverneur GN est un modificateur et si le GN n'est pas la fille d'un GN, le nom reçoit le trait *diloc=yes.*

Le transfert

Pour le transfert, nous n'avons généralement besoin que de règles de transfert simples. Pour traduire, par exemple, la structure d'arbre de la figure 3 en hollandais, les règles suivantes suffisent :

```
t1= {gb_lu=DILOC} => {nl_lu=DILOC}.
t2= {gb_lu=moment} => {nl_lu=moment}.
```

Dans quelques cas cependant, la traduction des GP temporels provoque des changements structuraux. Nous avons vu que certains noms temporels ont un article défini dans une langue, mais pas dans l'autre. Si nous reprenons l'exemple *on Monday / el lunes,* la traduction de l'anglais en espagnol peut être traitée avec une règle complexe qui peut s'appliquer à tous les jours de la semaine :

```
t3 = ~:{cat=pp}
     [~:{cat=p},
     NP:{cat=np}
          [N:{cat=n,diloc=yes,msdefs=msindef,gb_lu=monday; tuesday;
          wednesday; thursday; friday; saturday; sunday},
               REST:*{}]]
=> NP < N:{cat=n, msdefs=msdef}, REST >.
```

Cette règle efface le GP et le nœud de la préposition. Le GN est conservé dans la traduction, mais la valeur du trait du déterminant (*msindef*) est changée en *msdef.* Pour la traduction de l'espagnol en anglais, il suffit d'appliquer la règle inverse.

D'autres problèmes similaires se posent pour les heures et les parties du jour (le matin, etc.) qui demandent le génitif en hollandais.

at 3 o'clock = a las tres
in the morning = 's morgens

La génération

Pour la génération (hollandaise, dans cet exemple), nous avons besoin d'une règle qui identifie les GP qui contiennent le trait *diloc=yes* et qui insère la préposition correcte :

```
ts_diloc = PP:{cat=pp}
     [P:{cat=p,nl_lu=DILOC},
     NP:{cat=np}
          [N:{cat=n,diloc=yes,loc_temp_in=PREP},
          REST:* {}]]
=> PP < P:{nl_lu=PREP}, NP < N, REST> >.
```

Ainsi, la méthode lexicale garantit la traduction correcte de toutes les locatives directes.

Remarques finales

Dans cet article, nous avons essayé de montrer que les prépositions les plus problématiques pour la traduction peuvent être groupées dans une classe spéciale : celle des *locatives directes.* Cette classe de prépositions est mieux traitée par une méthode lexicale, basée sur les fonctions lexicales de Mel'čuk. Cette méthode permet un choix plus contrôlé de la traduction de prépositions temporelles et produit ainsi de meilleures traductions. La méthode paraît aussi prometteuse pour le traitement des prépo-

sitions spatiales. Des recherches ultérieures sont cependant nécessaires pour les GP's quantifiés (par exemple, *on the third morning*) et les noms composés (par exemple, *on Saturday morning*).

Références

BRÉE, D. S., SMIT, R. A. et J. P. VAN WERKHOVEN (1990) : « Translating Temporal Prepositions Between Dutch and English », *Jounal of Semantics*, 7, pp. 1-51.

BRÉE, D. S. (1992) : « Words for Time », V. Pouthas, W. Friedman et F. Macar (dir), *Time, Action and Cognition*, Kluwer, Dordrecht, pp. 337-348.

MEIJER, Siety, RIJKHOEK, Paulien, VANGILBERGEN, Ludo et Jesus VIDAL (1992) : *Interim Report for the CRC on Tense and Aspect*, Leuven, EUROTRA-NL/B, June 1992.

MEIJER, Siety, RIJKHOEK, Paulien, VANGILBERGEN, Ludo et Jesus VIDAL (1993) : *Final Report for the CRC on Tense and Aspect*, Leuven, EUROTRA-NL/B, January 1993.

MEL'ČUK, I. A. et A. K. ŽOLKOVSKY (1988) : « The Explanatory Combinatorial Dictionary », M. Evens (dir), *Relational Models of the Lexicon*, Cambridge University Press.

17

Pour une méthodologie de l'évaluation de la TA

Adriane RINSCHE

The Language Technology Centre, Kingston, Angleterre

• *Abstract* •

This paper is based on extensive studies of MT evaluation methods and practical comparative evaluation experience testing major MT systems such as SYSTRAN, METAL, LOGOS, ARIANE. Evaluation studies were carried out within the framework of a Ph.D, on behalf of the EC Commission, and the London based computer consultancy OVUM. The paper begins with some general observations on the state of the art of NLP and MT evaluation, followed by a presentation of evaluation methodologies for comparing the quality of the linguistic performance of either subsequent versions of one machine translation systems (vertical evaluation) or of several translation systems (horizontal evaluation). Merits and shortcomings of a rudimentary corpus based vertical evaluation methodology are briefly discussed. The horizontal evaluation method introduced consists of a combination of test suites and text samples. Translation quality is measured quantitatively by counting error frequencies. Some recommendations are also made regarding comparative linguistic performance testing of translation workbenches, particularly with regard to linguistic versus statistical fuzzy matching efficiency.

La situation actuelle

Pendant les 30 dernières années, de nombreux essais ont été faits pour développer des critères d'évaluation en traduction automatique. Cependant, jusqu'à présent, il n'existe pas encore de procédé généralement accepté. C'est pourquoi les fabricants, les chercheurs, les sponsors et les utilisateurs de systèmes doivent créer leurs propres critères et techniques d'évaluation, ou doivent utiliser des méthodes destinées à un système de traduction automatique qui ne peuvent pas être facilement transférées à d'autres systèmes ou applications.

Initiatives européennes de l'évaluation du TALN (traitement automatique des langues naturelles)

Pour remédier à cette situation, plusieurs initiatives européennes ont été prises récemment.

LISA (*Localization Industry Standards Association*), une organisation à but non lucratif dont l'objectif est de proposer des méthodologies et standards pour garantir que la création du logiciel multilingue, de la documentation technique et de la production multimédiatique soit d'un niveau général de qualité aussi élevé que possible et tienne compte des besoins des utilisateurs.

EAGLES (*Expert Advisory Group on Language Engineering Standards*) a été créé dans le cadre du programme *Linguistic Research and Engineering* (LRE) de la Commission des Communautés européennes. Le groupe est formé d'experts européens travaillant en vue de spécifications et directives communes dans les domaines du TALN. Un des cinq groupes de EAGLES se consacre à l'évaluation et au classement des systèmes de TA.

Une des activités importantes de EAMT, l'Association européenne de TA, est l'évaluation de la TA.

Dans le groupe allemand des utilisateurs de TA, un groupe d'évaluation, qui est constitué d'utilisateurs actuels et potentiels, a été mis sur pied de manière informelle.

L'importance de l'évaluation de la TA

Ces initiatives témoignent de l'importance du sujet. L'évaluation est une composante clé dans chaque technologie. Il est nécessaire de classer la performance générale des systèmes de TA. On peut comparer la performance linguistique d'un système aux différents niveaux de son développement. On nommera *vertical* ce type d'évaluation comparative. On peut, d'autre part, comparer la qualité de plusieurs systèmes. On nommera *horizontal* ce deuxième type d'évaluation.

L'évaluation verticale est importante pour les fabricants des produits de TALN et pour les vendeurs de services de traduction.

L'évaluation horizontale concerne les fabricants et les utilisateurs : les premiers pourront déterminer la qualité relative de leurs systèmes par comparaison aux produits de leurs compétiteurs ; les utilisateurs en puissance pourront décider en connaissance de cause avant d'acheter le produit qui satisfera le mieux leurs besoins.

Les constituants principaux de l'évaluation

Le classement de la qualité des systèmes de TA s'établit à partir des paramètres suivants :

– évaluation computationnelle et informatique ;

– facilité d'utilisation (*user friendliness*) ;
– évaluation de la performance linguistique et traductionnelle ;
– évaluation des implications organisatrices ;
– évaluation économique.

Chaque évaluation est partiellement déterminée par l'application envisagée.

Les exigences pour une méthodologie d'évaluation

Une méthodologie d'évaluation doit aboutir à des procédés standardisés internationaux pour mesurer la qualité de traduction.

- La méthodologie doit être pragmatique et transparente pour les utilisateurs.
- Il faut arriver à quantifier les jugements de la qualité des systèmes de TA.
- Les résultats doivent être obtenus dans un délai et à un coût raisonnables.

Les résultats des évaluations devraient influencer le développement d'un logiciel plus dirigé vers les besoins des utilisateurs.

L'évaluation de la performance linguistique des systèmes de TA

La méthodologie introduite ci-dessous est précédée d'une étude des méthodes d'évaluation utilisées dans le passé.

L'évaluation verticale

Exemple I : Systran

Une mesure simple de l'amélioration de la qualité de traduction de la version n+1 par rapport à celle de la version n du système de traduction Systran à la Commission des Communautés européennes a été développée récemment. Des contrôles préliminaires ont été effectués pour chacun des 16 couples de langues utilisés à la Commission.

On a créé un corpus de référence basé sur 3 000 phrases dans les deux langues sources, le français et l'anglais, 1 600 phrases sources en allemand et 2 500 phrases sources en espagnol. L'échantillonnage tient compte de la langue cible demandée, afin d'obtenir une image de la demande réelle transitant par le serveur Systran. Pour les 16 couples de langues, un corpus d'évaluation préliminaire de 437 000 phrases traduites automatiquement est créé.

Avec un logiciel simple, toutes les phrases traduites par la version actuelle de Systran (version n+1) sont comparées avec toutes les phrases traduites par la version précédente (version n). Toutes les phrases avec des traductions identiques dans n et n+1 sont éliminées, et toutes les traductions différentes (entre 3% et 12% du corpus) sont retenues comme corpus d'évaluation final. Toutes les traductions n et n+1 sont évaluées phrase par phrase. Elles sont présentées aux évaluateurs avec la phrase source et dans un contexte de deux ou trois phrases. Les évaluateurs sont des tra-

ducteurs professionnels. L'évaluation dure peu de temps : environ 1 minute par phrase. Cette méthode simple permet une décision rapide à l'égard de l'introduction de la nouvelle version de Systran. Elle est rudimentaire, parce que l'appréciation de l'évaluateur consiste à déterminer les niveaux d'amélioration ou de détérioration :

++ amélioration notable,
+ amélioration,
0 changement sans amélioration ou détérioration,
- détérioration,
-- détérioration notable.

Le pourcentage des phrases améliorées et détériorées est un indicateur de la qualité très approximatif mais qui pourrait gagner en précision, lorsqu'il aura été expérimenté davantage.

Exemple II : KANT : TA basée sur l'intelligence artificielle et dans des domaines restreints

Le système de TA KANT a été développé pour les produits de Caterpillar (tracteur à chenilles). L'objectif est de garantir que les utilisateurs puissent suivre les instructions des manuels traduits correctement. Ainsi, la stratégie d'évaluation en cours de préparation pour KANT se concentrera sur la justesse sémantique et la conformité du style de la TA. Les phrases seront réparties en deux ensembles : les phrases correctes et les phrases incorrectes. Le nombre de phrases correctes et incorrectes sera compté.

Les phrases incorrectes seront analysées de la manière suivante :

- les erreurs seront classées comme lexicales, morphologiques, syntactiques et sémantiques (voir ci-dessous, contrôle de performance linguistique) ;
- les erreurs sémantiques seront sous-classées en vertu de l'unité syntactique affectée (phrase nominale, phrase verbale, phrase prépositionnelle, etc.). Pour les erreurs qui seront uniquement identifiées dans un contexte plus large que celui de la phrase, la complexité du contexte nécessaire pour identifier et corriger l'erreur sera quantifiée ;
- en coopération avec les fabricants, la gravité des erreurs sera déterminée en fonction du temps nécessaire pour corriger l'erreur dans le système.

Pour les phrases correctes :

- les différences systématiques entre la TA et TH seront identifiées, classées et mesurées ;
- des catégories stylistiques seront développées au cours de l'évaluation.

Ces deux procédés de l'évaluation verticale ne peuvent pas être transférés sans modifications à l'évaluation horizontale. La méthode d'évaluation avec Systran n'est en aucun cas possible quand on compare des systèmes de traduction différents ; le nombre de phrases traduites de manière identique doit être négligeable pour les raisons suivantes :

- il y a pour chaque phrase source un nombre d'équivalents de traduction et

- il y a un même spectre plus grand de sources d'erreurs dépendant à la fois du système de TA et de la composition du dictionnaire.

La méthode d'évaluation à utiliser avec KANT intègre l'analyse des erreurs comme dans la méthode d'évaluation horizontale décrite ci-dessous. Une mesure quantitative et comparable du potentiel de plusieurs systèmes, c'est-à-dire la mesure du temps nécessaire pour corriger les erreurs dans le logiciel est probablement trop difficile et trop chère à utiliser dans une évaluation comparative.

Il faut, pour ces raisons, modifier les méthodes dans un contexte d'évaluation horizontale.

L'évaluation horizontale

Vue d'ensemble initiale des produits de TA

Pour chaque système et *workbench* de TA à comparer, les spécifications du système et ses caractéristiques seront décrites sous la forme d'un quadrillage afin de permettre une vue d'ensemble initiale du produit.

Il est nécessaire de couvrir les informations suivantes :

nom du système ;
nom du propriétaire ;
nom du distributeur ;
adresse ;
personne ressource.

Paires de langues traitées.

Dictionnaires :
types de(s) dictionnaire(s) (monolingue, bilingue, multilingue, réversible) ;
nombre d'entrées lexicales ;
structure des entrées lexicales.

Type de système (direct/transfert/interlingue).

Prix de vente :
achat du système de base ;
coût de livraison et d'installation ;
coût des contrats d'entretien ;
coût d'une licence annuelle pour le logiciel de TA ;
coût par paire de langues ;
coût par mot traduit.

Prix d'exploitation :
coût d'opération ;
coût de(s) dictionnaire(s) ;
coût de la formation des usagers ;

coût de la révision humaine.

Spécification :
marque de l'ordinateur principal ;
système d'exploitation ;
type de CPU ;
espace mémoire principale disponible ;
espace mémoire principale utilisée ;
disques durs ;
console et station de travail ;
éditeur (*editor*) ;
traitement (*Processing*) : *Interactive/Batch* ;
compatibilité réseau ;
record formats ;
filtres ;
protection du format (source/cible) ;
mémoire de traduction (*Identical/Linguistic/Statistical Fuzzy Match*) ;
facilités d'administration des banques de données (*Database Management Facilities*) ;
modèles grammaticaux ;
sémantique.

Contrôle des Performances Linguistiques

Les produits *workbench*

On appellera produit *workbench* tout système de TA comprenant au moins une mémoire de traduction et une facilité pour reconnaître les phrases sources identiques (*matching*) et similaires (*fuzzy matching*). L'efficacité de la reconnaissance et du remplacement des phrases similaires (*fuzzy matching*) sera contrôlée. Reste à développer une procédure pour comparer le *fuzzy matching*. La méthode doit être basée sur un corpus et tenir compte d'au moins deux aspects :

- l'administration des ensembles des données et l'efficacité de *matching* des phrases similaires :
  - l'identification des phrases sources similaires divisées en deux parties ou plus dans les versions futures du manuel d'utilisation ;
  - l'identification des phrases sources similaires, quand deux phrases ou plus sont combinées dans les versions futures du manuel d'utilisation ;
  - le niveau de complexité jusqu'où les différences des textes sources sont administrées de manière satisfaisante par le système ;
- assistance pendant la traduction : l'efficacité de *matching* des mots et expressions :
  - le niveau maximal de dissimilarité des ensembles de données auquel un système peut encore identifier des sous-structures identiques.

Une simple méthode préliminaire et objective pourrait inclure au moins le calcul des *fuzzy matches* réussis et le nombre des matches potentiels non reconnus par le système. De ces résultats, pourrait être déduit un profil de l'efficacité du *matching*.

Pour les systèmes de TA, une technique plus sophistiquée a été développée :

dans chaque cycle de contrôle deux types de test seront utilisés :
1. batterie de tests (*Test Suites*) ;
2. échantillons de textes.

Une batterie de tests est un ensemble de phrases qui permet de contrôler les phénomènes grammaticaux couverts par le système. Il devrait être global et indépendant du contexte donné. Dans un contexte comparatif il permet de juger du degré de sophistication grammaticale des systèmes.

On utilise des échantillons pour contrôler l'aptitude d'un système pour un domaine d'application spécifique.

Il faut passer les contrôles dans l'ordre suivant :

1re étape :	échantillons : traduction brute sans mise à jour du dictionnaire, mesure vitesse de traduction ;
2e étape :	mise à jour du dictionnaire, mesure vitesse et efficacité ;
3e étape :	échantillons : traduire de nouveau, batterie de tests : traduire, mesure vitesse de traduction ;
4e étape :	évaluation des traductions obtenues dans les 1re et 3e étapes en comptant : a) le nombre de phrases sans erreur et b) le nombre et les types d'erreurs dans les deux types de tests ;
5e étape :	interprétation et comparaison des statistiques des erreurs.

Il faut introduire une procédure pour peser les erreurs suivant leur gravité. La gravité des erreurs pourrait être déterminée de manière pragmatique suivant leurs implications pour la postédition. Une méthode plus systématique serait de déterminer la gravité des erreurs suivant le temps nécessaire pour corriger l'erreur dans le logiciel. Ces informations sont cependant difficiles à obtenir et varient d'un système à l'autre dans un contexte d'évaluation horizontale.

L'analyse des erreurs

La qualité linguistique des deux types de tests est analysée en attribuant toutes les erreurs à un nombre de catégories pour déterminer la concentration des erreurs dépendant du système dans chacune des catégories suivantes :

Catégorie I : Les erreurs lexicales

Une erreur lexicale est attribuée quand une unité lexicale n'est pas du tout ou incorrectement traduite. Comme sous-classe, les erreurs qui résultent du choix d'un verbe sont classées séparément, parce que la fréquence des erreurs de verbes est particulièrement indicative de la qualité linguistique du système à cause des conséquences syntactiques et sémantiques possibles d'une traduction incorrecte ou de la non-traduction du verbe. On distingue les catégories suivantes :

unité lexicale	– non-traduction
unité lexicale	– traduction incorrecte
verbe	– non-traduction
verbe	– traduction incorrecte

Catégorie II : Les erreurs morphologiques

Il n'y a pas encore d'expériences parce que l'anglais a toujours été la langue cible.

Catégorie III : Les erreurs syntactiques

Cette catégorie d'erreurs concerne les erreurs dans la génération des phrases sans conséquences sémantiques. On peut encore reconnaître le sens de la phrase, de la proposition ou de l'expression. Les sous-catégories suivantes sont distinguées :

structure syntactique de la phrase	– incorrecte
phrase verbale	– incorrecte
phrase nominale	– incorrecte
phrase prépositionnelle	– incorrecte
proposition subordonnée	– incorrecte

La première catégorie de cette classe d'erreurs est appliquée aux phrases qui présentent un désordre syntaxique général, mais dont le sens n'est pas affecté. Une phrase complexe qui consiste en plusieurs phrases simples coordonnées est divisée en ses sous-structures, et chacune de ces sous-structures compte comme une phrase.

Les erreurs syntactiques de la phrase verbale, nominale et prépositionnelle concernent la génération incorrecte de ces sous-structures. Les erreurs dans des phrases verbales, nominales et prépositionnelles coordonnées sont comptées séparément.

Les mêmes principes sont appliqués à l'évaluation des phrases subordonnées.

Catégorie IV : Les erreurs sémantiques

Les erreurs sémantiques concernent le sens de la phrase complète, ou le sens des phrases nominales, verbales et prépositionnelles. Les autres catégories se rapportent aux erreurs de référence et aux expressions idiomatiques :

sens de la phrase	– incorrecte
sens de la phrase nominale	– incorrecte
sens de la phrase verbale	– incorrecte
sens de la phrase prépositionnelle	– incorrecte
expression idiomatique	– incorrecte
erreur de référence	

Les phrases ou sous-phrases coordonnées sont comptées et analysées séparément comme dans catégorie III.

Catégorie V : Les erreurs dans le texte source

Les erreurs dans le texte source ne devraient pas exister. Elles apparaissent éventuellement pour les raisons suivantes :

- fautes de frappe dans le texte matériel ou introduites au cours du transfert sous forme électronique ;
- ambiguïtés du texte source qui ne peuvent pas être résolues par le logiciel.

Il faut les compter de manière isolée pour faire une évaluation correcte du système.

L'évaluation et l'interprétation de la qualité de la TA

Venons-en aux résultats de l'évaluation de la qualité de la traduction. On distingue :

1. qualité de la traduction brute (échantillons) basée sur le calcul des erreurs (quantitative) ;
2. qualité de la souplesse des dictionnaires par analyse :
   (a) de la facilité d'utilisation (*user friendliness*) (descriptive),
   (b) l'organisation linguistique (descriptive),
   (c) de la vitesse de mise à jour (*update*) (quantitative),
   (d) erreurs pendant et après la mise à jour (*update*) (quantitative) ;
3. qualité de la traduction après la mise à jour pour :
   – les batteries de tests (*test suites*) et
   – les échantillons de textes
   basée sur le calcul des erreurs, avec des résultats intéressants dépendant du type de test et du système qui mène à :
   – un profil de la performance grammaticale,
   – un profil à trois dimensions de l'aptitude (système/couple de langue/domaine de sujet) ;
4. qualité générale par système contrôlé :
   – interprétation des résultats (descriptive) ;
5. comparaison des systèmes (descriptive et quantitative).

Des tests antérieurs ont montré que la concentration des erreurs varie considérablement selon le type de tests et le système contrôlé. La fréquence des erreurs dans les batteries de tests témoignent en général de la qualité du champ grammatical ; les fréquences des erreurs dans les échantillons sont révélatrices de l'aptitude d'un sous-ensemble de caractéristiques grammaticales pour une application spécifique.

Dans un petit nombre de cas, un certain degré de subjectivité subsiste peut-être dans le processus d'attribution des erreurs aux catégories, parce que l'analyse de sortie *boîte noire* peut créer des problèmes de détermination correcte des causes d'erreurs. Les comparaisons de la qualité de TA basées de manière consistante sur l'analyse des phénomènes de surface via l'analyse d'erreurs sont cependant la méthode la plus objective pour mesurer la performance linguistique. Des comparaisons des fréquences d'erreurs dans différents systèmes sont basées sur les résultats quantitatifs et constituent donc une base adéquate et valable pour des mesures de qualité comparatives et quantitatives. Cette procédure est le seul moyen objectif et quantifiable de mesure directe de la qualité de la traduction. Le fait que d'importants fabricants de logiciels tels que DIGITAL mesurent la qualité de la traduction humaine en termes de

fréquences d'erreurs confirme d'autant plus la méthode choisie. Les utilisateurs de TA pourront naturellement adopter la même stratégie pour évaluer leurs résultats.

Conclusion

L'évaluation de logiciels de TALN est une discipline jeune qui n'a pas encore suffisamment mûri. Beaucoup de travail reste à faire avant que des méthodes et des résultats satisfaisants et acceptés à l'échelle internationale soient disponibles. La nature de la langue naturelle même empêche des méthodes d'évaluation complètement objectives et formelles. Quelle que soit la qualité quantitative et automatique d'un procédé, un certain élément de subjectivité persistera. Il est nécessaire de minimiser cet élément et la quantité d'effort humain dans l'évaluation pour arriver à des solutions utilisables et financièrement viables. Il est à espérer que les procédés décrits ci-dessus stimuleront les débats à venir.

Références

BOURBEAU, Laurent (1990) : *Élaboration et mise au point d'une méthodologie d'évaluation linguistique de systèmes de traductions assistées par ordinateur,* Rapport final, Secrétariat d'État du Canada.

FALKEDAL, Kirsten (sd) : *A Practical Guide to the Evaluation of Machine Translation Systems,* ISSCO, Interim Report to Suissetra.

RINSCHE, Adriane (1993a) : *Evaluationsverfahren für maschinelle Übersetzungssysteme. Zur Methodik und experimentellen Praxis,* Thèse, Université de Bonn, Publié, Kommission der Europäischen Gemeinschaften, Informationsmanagement, EUR 14766 DE. ISSN 1018-5593.

RINSCHE, Adriane (1993b) : « The Rinsche MT Evaluation Methodology – REM », *The Language Industry Tribune,* Édition spéciale: Industrial Convention « Linguistic Engineering », Paris, OFIL.

RINSCHE, Adriane (1993c) : « Towards a MT Evaluation Methologogy », *Proceedings. The Fifth International Conference on Theoretical and Methodological Issues of Machine Translation,* Kyoto, Japon.

RINSCHE, Adriane (1993d) : « Towards a User-oriented Evaluation Methodology for Computer Translation Tools », *The LISA Forum Newsletter,* vol. 2, pp. 17-20.

RINSCHE, Adriane (1991i) : *Towards a System of Benchmarking MT Systems,* Report for EC Commission, October 10, 1991.

18

Structure communicative de l'énoncé dans la génération automatique du texte

Lidija IORDANSKAJA

Université de Montréal, Montréal, Canada

• *Abstract* •

A text generation model presented here essentially uses the communicative structure of an utterance, i.e. *information about 1) new vs given, 2) theme vs rheme, and 3) emphasis. In our model (which is based on Meaning-Text Theory) the synthesis of a sentence consists of series of transitions between representations of different levels: from the semantic representation to the deep-syntactic one, then to the the surface-syntactic representation, then to the morphological one, then either to the graphic representation (written text), or to the phonological, phonetic and acoustic representations (speech). The communicative structure is used during the three first transitions of the synthesis of written text and during the transition to the phonological representation – for determinating the prosody of a sentence.*

Introduction

Nous présentons ici un modèle de génération de texte qui est sous-jacent aux trois systèmes de génération des rapports statistiques – les rapports sur le marché de travail (LFS – *Labour Force Statistics*), sur les ventes de détail (RTS – *Retail Trade Statistics*) et sur l'indice des prix à la consommation (CPIS – *Consumer Price Index Statistics* ; rapports anglais et français). Ces systèmes ont été élaborés par une équipe comprenant M. Kim, R. Kittredge, B. Lavoie, A. Polguère et L. Iordanskaja (voir Iordanskaja *et al.* 1992). Des tables statistiques constituent l'entrée des systèmes et leur sortie est un texte *verbalisant* les tables données.

Comme tout système de génération de texte, les systèmes en question comprennent un composant planificateur et un composant linguistique.

Le *composant planificateur* effectue les trois tâches suivantes :

- il détermine le contenu du texte à générer, en le présentant comme un ensemble des représentations conceptuelles, chaque représentation correspondant à un message ;
- il détermine la structure du texte, en ordonnant les messages et en les distribuant en énoncés (phrases futures) ;
- il effectue le passage de la séquence des représentations conceptuelles des messages à une séquence des représentations sémantiques des énoncés, qui constitue l'entrée pour le composant linguistique.

Comme on le voit, nous distinguons la représentation conceptuelle du texte et sa représentation sémantique : la première est orientée vers le domaine décrit et ne dépend pas de la langue de sortie (ce qui rend possible de l'utiliser comme une *interlingua* dans la génération multilingue) ; la deuxième, par contre, est adaptée à la langue de sortie (le lecteur trouvera des exemples des représentations conceptuelles et sémantiques plus loin).

Le *composant linguistique* effectue la synthèse des phrases à partir des représentations sémantiques. Le composant linguistique s'appuie sur le modèle linguistique Sens-Texte (voir, par exemple, Mel'čuk 1978), où la synthèse d'une phrase se déroule comme une série de passages entre représentations linguistiques des différents niveaux : sémantique, syntaxique profond, syntaxique de surface, morphologique et graphique. Au cours de la synthèse d'une phrase, le composant linguistique utilise de façon essentielle la structure communicative de l'énoncé ; cette utilisation constitue la cible centrale du présent article.

Le texte qui suit se divise en deux parties : dans la prochaine section, nous présentons les notions communicatives exploitées et décrivons pour quel but et comment le composant linguistique les met en jeu. La dernière section est consacrée au problème du calcul de la structure communicative dans la représentation sémantique, problème relevant du composant planificateur.

La structure communicative d'un énoncé et la synthèse d'une phrase

Notions utilisées

De tous les aspects de la structure communicative d'un énoncé, nous ne mettons en jeu que deux types d'information :

1) la structure thématique de l'énoncé ;
2) la mise en relief de certains éléments de l'énoncé.

Structure thématique

Thème et rhème

La structure thématique implique deux notions principales : le thème et le rhème. Le THÈME (T) d'un message est ce dont le message parle, ce à quoi il est consacré. Le RHÈME (R) d'un message est ce qui est dit à propos du thème.

Dans un réseau sémantique, le thème et le rhème peuvent contenir plusieurs éléments, dont l'un est spécifié comme dominant. L'élément DOMINANT du thème (ou du rhème) est un élément que les autres éléments du thème (ou du rhème) modifient.

La phrase :

(1) *L'emploi chez les femmes a diminué de 0.1%.*

a dans sa représentation sémantique (RSém), présentée dans la figure 1, le thème « emploi --> femmes », avec le nœud dominant « emploi » ; et le rhème « diminuer --> 0.1% », avec le nœud dominant « diminuer ».

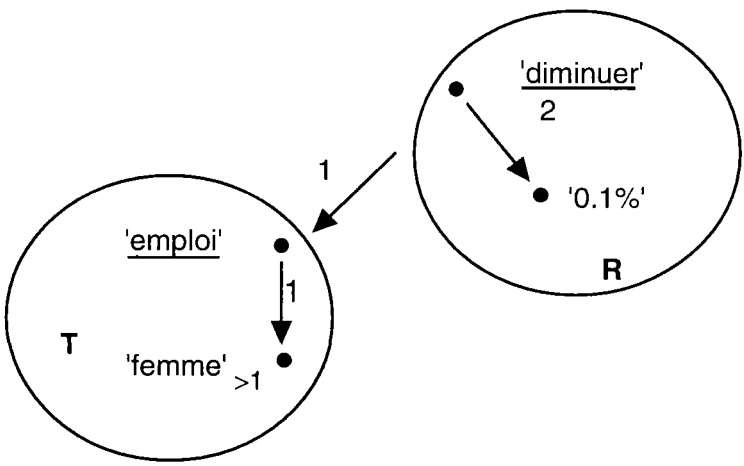


FIGURE 1 : RSém de la phrase (1).

Spécificateur

Un énoncé peut contenir des éléments qui n'appartiennent ni au thème ni au rhème. Tels sont les connecteurs, les expressions parenthétiques et les éléments circonstanciels autonomes, qui spécifient la situation décrite dans sa totalité. Nous appelons ces éléments SPÉCIFICATEURS (SP). Voir (2)-(4) :

(2) ***D'un mois à l'autre*** (SP), *l'indice d'ensemble* (T) *a régressé de 0.2%* (R).
(3) ***Après désaisonnalisation*** (SP), *l'indice d'ensemble* (T) *a avancé de 0.3% en octobre* (R).

(4) ***En septembre*** (SP), *l'indice d'ensemble* (T) *n'a pratiquement pas changé* (R).

Dans ces phrases, les syntagmes prépositionnels initiaux (= des SP) ne font partie ni du thème (qui est *l'indice d'ensemble*) ni du rhème. Un tel syntagme est un spécificateur de toute la situation décrite. Voir (5), où un syntagme du même type caractérise l'indice et non pas la situation entière ; il fait partie du thème :

(5) *L'indice d'ensemble en septembre n'a pratiquement pas changé.*

Mise en relief

Le locuteur peut METTRE EN RELIEF certains éléments de la phrase, pour y attirer l'attention spéciale du destinataire. C'est le fait notamment quand ces éléments entrent dans des oppositions importantes. Par exemple, dans (2) et (3), les syntagmes prépositionnels initiaux sont mis en relief, parce que leurs oppositions sont importantes pour le rapport en question. La phrase (2) est la première phrase du paragraphe consacré aux changements mensuels, à la différence des changements annuels ; la phrase (3) commence un paragraphe qui est consacré aux estimations de l'indice désaisonnalisé des prix à la consommation, tandis que, dans les paragraphes précédents, l'indice était **non** désaisonnalisé.

Utilisation de la structure communicative pendant la synthèse d'une phrase

Utilisation de la structure thématique

En effectuant la synthèse d'une phrase, le composant linguistique utilise la structure thématique pendant le passage de la représentation sémantique (RSém) à la représentation syntaxique profonde (RSyntP) – pour déterminer :

1) le choix du sommet de l'arbre syntaxique profond ;
2) le choix entre actif et passif ;
3) le choix de certaines paraphrases impliquant des fonctions lexicales.

Choix du sommet de l'arbre syntaxique profond

La structure sémantique – le composant principal de la RSém – est un réseau et la structure syntaxique profonde – le composant principal de la RSyntP – est un arbre, qui, à la différence du réseau, a un sommet. Comme l'a montré Polguère (1990), la structure communicative joue un rôle majeur dans le processus de passage d'un réseau sémantique à l'arbre syntaxique. Pour passer du réseau à l'arbre, il faut déterminer dans le réseau un nœud qui donnera le sommet de l'arbre correspondant. Les règles qui déterminent ce nœud sommet, s'appuient sur la structure thématique. Ainsi, la première option est de choisir, en tant que la source pour le sommet, le nœud dominant du rhème. La phrase (6) :

(6) *Cette hausse poursuit la tendance amorcée en avril.*

a la RSém présentée dans la figure 2 et la RSyntP présentée dans la figure 3.

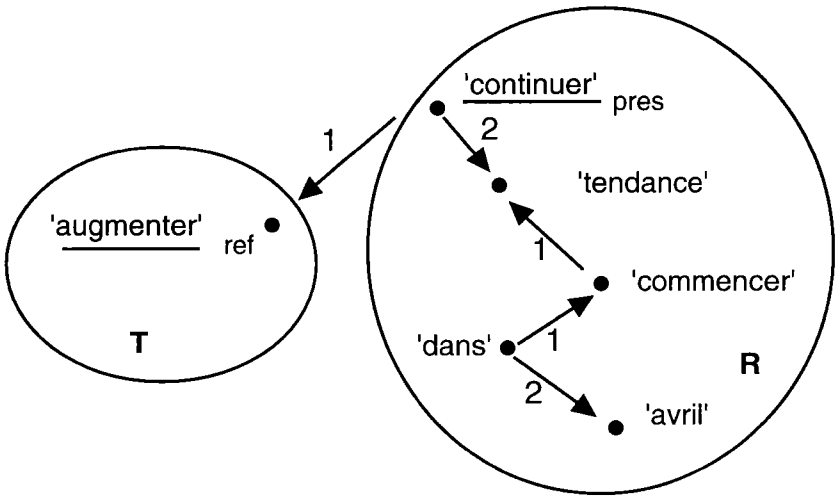


FIGURE 2 : RSém de la phrase (6).

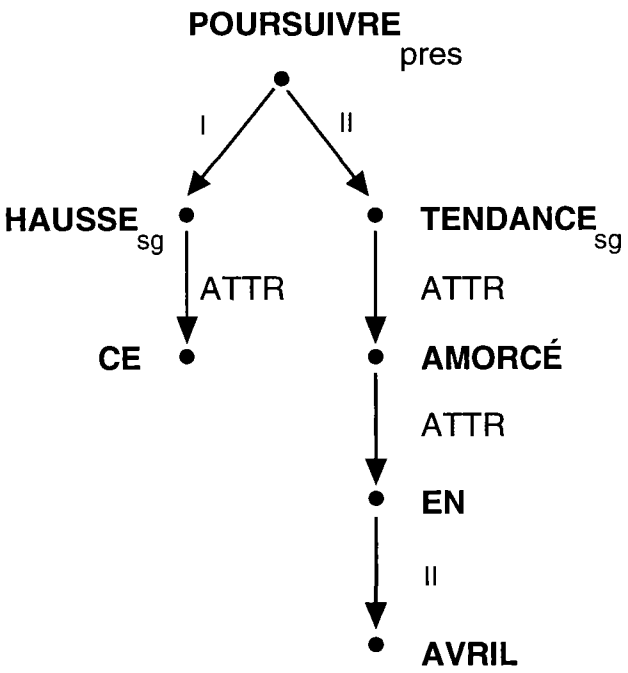


FIGURE 3 : RSyntP de la phrase (6).

Nos règles choisissent, comme nœud sommet, le nœud « continuer » (qui donnera plus tard *poursuivre*) parce que « continuer » est le nœud dominant du rhème. Par exemple, le nœud « commencer » (qui donne *amorcer*), ne peut pas être le sommet de l'arbre, bien qu'il entre dans le rhème, parce qu'il n'est pas son nœud dominant.

Choix entre actif et passif

Le composant linguistique utilise la règle suivante : si dans la RSém réduite le thème n'est pas le premier actant du verbe sommet, le grammème *passif* doit être attribué à ce verbe (bien entendu, si celui-ci admet la passivation).

La RSém réduite (RSémR) est une représentation intermédiaire entre les niveaux sémantique et syntaxique profond, utile du point de vue procédural ; les nœuds du réseau sémantique réduit sont des sens correspondant aux lexèmes futurs et un nœud du réseau réduit est spécifié comme sommet de l'arbre futur.

La phrase :

(7) *Ces baisses ont été compensées par les hausses des ventes des chaussures.*

a la RSémR, présentée dans la figure 4. Dans cette RSémR, le sommet de l'arbre futur est « compenser », dont le premier actant est « augmenter » (*les hausses des ventes des chaussures*) et le deuxième actant est « diminuer » (*ces baisses*). Le thème « diminuer » n'est pas le premier actant de « compenser » ; par conséquent, la construction passive est choisie en (7).

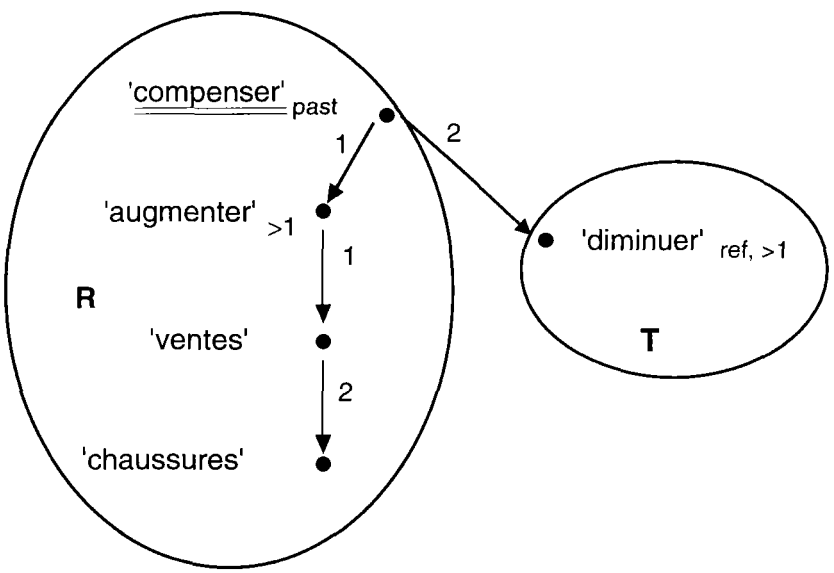


FIGURE 4 : RSémR de la phrase (7).

Choix des paraphrases impliquant des fonctions lexicales (FL)

La notion de la FL, proposée dans la théorie Sens-Texte, la liste des FL et la façon de les utiliser dans les dictionnaires sont présentées dans l'article de I. Mel'čuk (dans cet ouvrage).

Le composant linguistique met en jeu quelques FL pour deux tâches :1) rendre compte de la cooccurrence lexicale restreinte (du type ***au*** *Québec* mais ***en*** *Ontario*) et 2) varier le texte de sortie en choisissant des paraphrases différentes (voir Mel'čuk 1988a, b ; 1992). Le composant linguistique possède quelques règles de choix entre paraphrases, qui s'appliquent pendant le passage de la RSémR à la RSyntP. Ces règles s'appuient sur la structure thématique de la RSémR. Donnons comme exemple trois règles de ce type.

1) La phrase (8b) est obtenue comme résultat d'application d'une règle de paraphrase présentée dans la figure 5.

(8) a. *Les ventes* (T) *ont diminué de 0.1% en octobre.*
b. *Les ventes ont affiché* [$Oper_1$(DIMINUTION)] *une diminution* [S_0(DIMINUER)] *de 0.1% en octobre.*

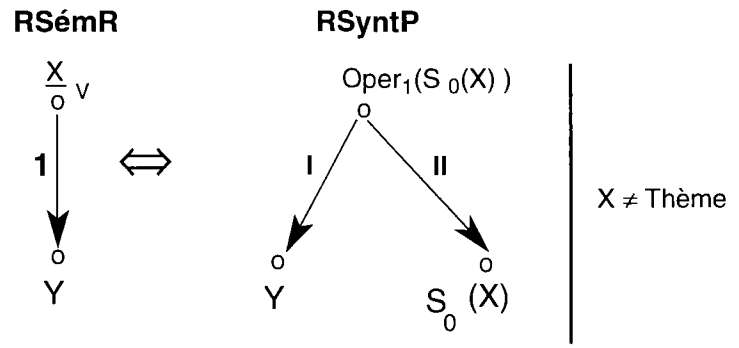


FIGURE 5 : Règle 1 de paraphrase.

Cette règle assure le remplacement d'un verbe X par un nom déverbatif correspondant S_0(X) et $Oper_1$ de ce nom ; elle peut être utilisée à la condition que le verbe à remplacer ne soit pas le thème. En (8a), X = « diminuer », et « diminuer » n'est pas le Thème ; donc, la règle indiquée peut être appliquée. Si dans la RSémR « diminuer » est le thème, cette RSémR ne peut pas être réalisée comme (1b) ; elle doit être réalisée comme (8c) :

(8) c. *Une diminution de 0.1%* (T) *a été enregistrée* [$Func_1$(DIMINUTION)] *pour les ventes en octobre.*

2) La phrase (9b) est obtenue comme résultat d'application d'une règle de paraphrase, présentée dans la figure 6.

(9) a. *Les ventes* (T) *ont diminué dans toutes les provinces.*
b. *Toutes les provinces* (T) *ont affiché* [$Oper_1$(BAISSE)] *des baisses de ventes.*

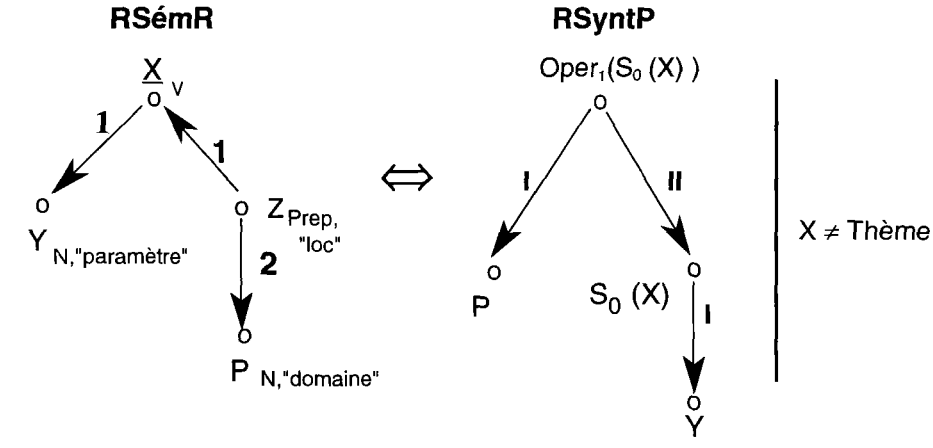


FIGURE 6 : Règle 2 de paraphrase.

Cette règle effectue une transformation d'une configuration contenant un verbe X, le nom du *paramètre* Y comme son premier actant et deux nœuds – Z et P – caractérisant la localisation de ce paramètre ; elle peut être appliquée sous la condition que P (le nom du *domaine*) soit le thème.

3) La phrase (10b) est obtenue comme résultat d'application d'une règle de paraphrase, présentée dans la figure 7, qui, comme on peut le voir, utilise, elle aussi, la structure thématique de la RSémR.

(10) a. *En septembre, le niveau de l'emploi est estimé à 11 747 000, c'est-à-dire ce niveau a augmenté de 31 000 par rapport à septembre.*
b. *En septembre, le niveau de l'emploi est estimé à 11 747 000, en hausse* [Adv_1(AUGMENTER)] *de 31 000 par rapport à septembre.*

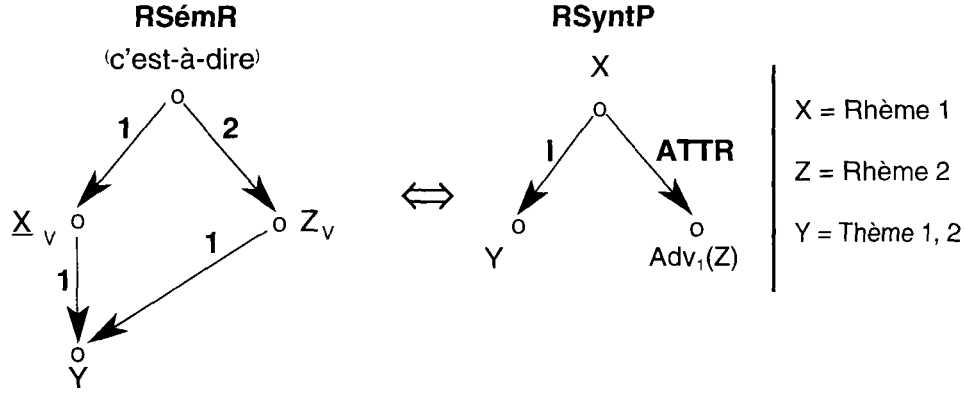


FIGURE 7 : Règle 3 de paraphrase.

Utilisation de la mise en relief

Nous venons de montrer comment la structure thématique est utilisée. Le composant linguistique met également en jeu un autre type d'information communicative – la mise en relief de certains éléments. Il s'agit, dans nos textes, des spécificateurs mis en relief, illustrés en (2)-(4).

Les spécificateurs mis en relief reçoivent une marque spéciale dans la RSém, et cette marque est utilisée pendant le passage de la représentation syntaxique de surface à la représentation morphologique pour deux tâches :

1. la détermination de l'ordre des mots (un spécificateur mis en relief doit être placé au début de la phrase) ;

2. la détermination de la ponctuation (un spécificateur mis en relief doit être suivi par une virgule).

Problème de calcul de la structure thématique d'un énoncé

Comme on l'a déjà dit, les représentations sémantiques sont obtenues à partir des représentations conceptuelles (RConc) des messages et elles héritent les indications du thème et du rhème de ces RConc. Les règles d'héritage sont maintenant banales : le transfert s'effectue SANS AUCUN CHANGEMENT de la structure thématique de la RConc donnée. Cependant, la structure thématique au niveau conceptuel n'est pas encore orientée vers le lecteur du texte futur : elle reflète le processus de recherche des informations pertinentes (le thème correspond à des informations données et le rhème, à des informations recherchées). Dans la RSém, la structure thématique doit être adaptée aux besoins du lecteur humain : le texte doit être cohérent et naturel. La séquence des structures thématiques des énoncés joue un rôle important pour assurer ces propriétés du texte. Il s'ensuit qu'un *bon* système de génération doit avoir des règles de transformation de la structure thématique *conceptuelle* en structure thématique *sémantique*. Nous voulons proposer ici les règles pour deux cas particuliers qui sont typiques des rapports statistiques. Ces règles illustrent bien la pertinence de l'aspect rhétorique du texte pour les structures thématiques de ses énoncés.

L'idée générale que la génération de texte doit tenir compte de l'aspect rhétorique des textes est plus ou moins admise. Pourtant les chercheurs discutent beaucoup COMMENT cet aspect doit être représenté et utilisé. Une approche assez populaire propose que le texte soit représenté par une *structure rhétorique* – un arbre où les relations nommées rhétoriques (d'une liste prédéfinie) relient des fragments du texte (voir, par exemple, Mann et Thompson 1987). Nous allons montrer qu'au moins deux relations rhétoriques – contraste et énumération – peuvent être utiles pour la détermination de la structure thématique d'un énoncé.

Une remarque importante : la détermination de la structure thématique d'un énoncé n'est possible qu'après l'ordonnancement des énoncés.

Nous ne considérons que les séquences des MESSAGES *COMPLÉMENTAIRES* – messages qui manifestent le même schéma, ou, pour utiliser la terminologie courante, qui cor-

respondent au même patron (angl. *template*). Ces messages décrivent le même type de phénomène mais pour des cas particuliers différents ; voir les exemples (11) et (12). Les séquences des messages complémentaires se divisent en séquences contrastives et séquences énumératives.

1) *Séquences contrastives*

Soit la séquence des deux messages suivants :

(11) 1. *L'emploi chez les hommes* (T) *a augmenté de 0.3%.*
2. *L'emploi chez les femmes* (T) *a diminué de 0.1%.*

Ces messages résultent du même patron du plan présenté dans la figure 8.

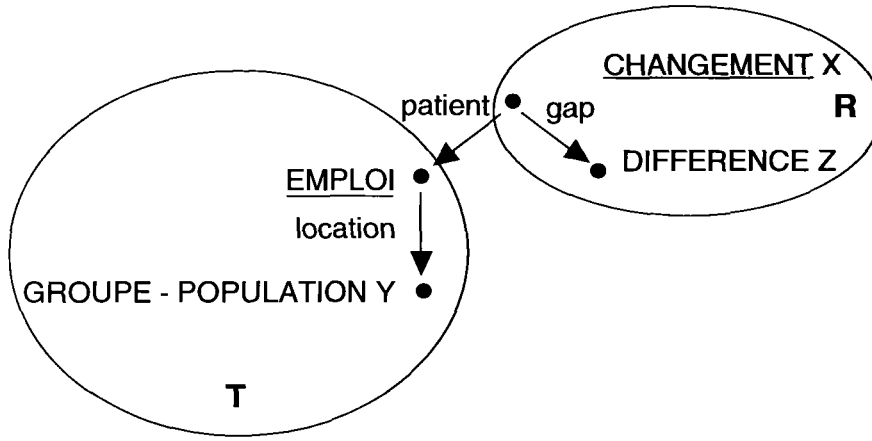


FIGURE 8 : Patron pour les messages du type (11).

Dès que le planificateur fait jouer les variables du patron, il obtient les RConc des messages, leur structure thématique héritant de la structure thématique spécifiée dans le patron ; voir la figure 9.

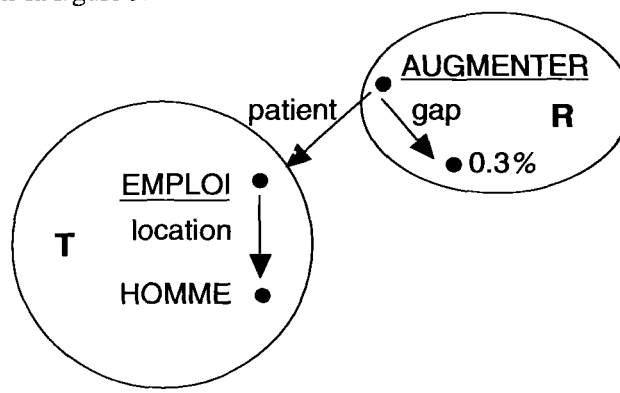


FIGURE 9 : RConc du message (11.1).

Les messages (11.1) et (11.2) sont des messages complémentaires. Puisque les messages complémentaires correspondent à un même patron, ils ont les structures thématiques conceptuelles identiques. Dans les deux RConc correspondant aux (11.1) et (11.2), le thème est « EMPLOI [concept dominant] pour un GROUPE » et le rhème est « CHANGEMENT [concept dominant] et DIFFERENCE ». Si on réalise ces structures thématiques directement, on obtient le texte (11), qui, bien sûr, n'est pas satisfaisant du point de vue stylistique – avant tout, parce que le contraste entre deux éléments (« emploi chez les hommes » vs « emploi chez les femmes ») n'est pas exprimé (dans le texte écrit où il n'y a pas de prosodie). Une façon possible de le rendre plus naturel est la suivante.

Le planificateur, dans le cas où le nombre de messages complémentaires est égal à DEUX, indique qu'entre les deux messages correspondants il y a une relation rhétorique *contraste.* Cette relation peut être exprimée soit par un connecteur correspondant (par exemple, *par contre* ou *alors que*), soit par la structure thématique du deuxième message. Si le générateur choisit la deuxième option, il doit transformer, en passant à la RSém, le deuxième élément contrastif (ici, « les femmes ») en élément dominant du thème ou, encore mieux, en élément dominant du thème mis en relief. Par conséquent, on obtient les textes (12) ou (13) :

(12) 1. *L'emploi chez les hommes* (T) *a augmenté de 0.3%.*
2. *Les femmes* (T) *ont affiché une diminution de 0.1%.*

(13) 1. *L'emploi chez les hommes* (T) *a augmenté de 0.3%.*
2. *Pour les femmes* (T_1), *l'emploi* (T_2) *a diminué de 0.1%.*

2) *Séquences énumératives*

Considérons une autre séquence des messages, où le nombre de messages complémentaires est supérieur à deux, ce qui est typique, par exemple, d'un paragraphe qui décrit les changements des ventes dans les dix provinces du Canada.

(14) 1. *Les ventes* (T) *ont augmenté de 0.5% au Québec et en Ontario.*
2. *Les ventes* (T) *ont peu augmenté en Colombie Britannique.*
3. *Les ventes* (T) *ont diminué de 0.3% au Manitoba.*
4. *Les ventes* (T) *n'ont pas changé dans les autres provinces.*

Rhétoriquement, cette séquence est une énumération de faits du même type. Tous les messages ont les RConc pareilles, avec la structure thématique identique : le thème = « VENTES [concept dominant] dans PROVINCE », le rhème = « CHANGEMENT [concept dominant] et DIFFÉRENCE ». Encore une fois, si on réalise ces structures thématiques directement, on obtient le texte (14), qui n'est pas naturel. Pour l'éviter, le planificateur doit produire, dans le cas où le nombre de messages complémentaires est supérieur à deux, une indication sur la relation *énumération* entre chacun des deux messages voisins.

Pour l'organisation communicative des textes énumératifs, une des stratégies possibles est la suivante :

1) dans le premier message, l'élément distinctif du thème conceptuel (ici – une

province) devient un élément non dominant du rhème sémantique (un cas typique des rapports statistiques composés par des rédacteurs humains) ;

2) le thème d'un message M_n est un élément qui, soit coïncide avec un des éléments du rhème du M_{n-1}, soit y est opposé.

Suivant cette stratégie, on obtient le texte (15), qui semble assez naturel :

(15) 1. *Les ventes* (T) *ont augmenté de 0.5% au Québec et en Ontario.*
2. *Une augmentation modeste* (T) *a été enregistrée en Colombie Britannique.*
3. *Manitoba* (T) *a affiché une diminution [des ventes] de 0.3%.*
4. *Les ventes* (T) *n'ont pas changé dans les autres provinces.*

Références

IORDANSKAJA, L. (1992) : « Communicative Structure and its Use during Text Generation », *International Forum on Information and Documentation,* vol. 17, n° 2, pp. 15-27.

IORDANSKAJA, L., M. KIM, R. I. KITTREDGE, B. LAVOIE et A. POLGUÈRE (1992) : « Generation of Extended Bilingual Statistical Reports », *Proceedings of 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, pp. 1019-1023.

IORDANSKAJA, L., M. KIM et A. POLGUÈRE (1992) : « Some Procedural Problems in the Implementation of Lexical Functions », K. Haenelt, L. Wanner (dir), « *International Workshop on the Meaning-Text Theory*, July 1992, pp. 197-205.

MANN, W. C. et S. THOMPSON (1987) : « Rhetorical Structure Theory: a Theory of Text Organization », *ISI Reprint Series,* ISI-RS-87-190, University of Southern California.

MEL'ČUK, I. A. (1978) : « Théorie du langage, théorie de la traduction », *META,* vol. 23, n° 4, pp. 271-302.

MEL'ČUK, I. A. (1988a) : « Paraphrase et lexique dans la théorie linguistique Sens-Texte », *Cahiers de Lexicologie*, vol. 52, n° 1, pp. 5-50.

MEL'ČUK, I. A. (1988b) : « Paraphrase et lexique dans la théorie linguistique Sens-Texte », *Cahiers de Lexicologie,* vol. 53, n° 2, pp. 5-53.

MEL'ČUK, I. A. *et al.* (1992) : *Dictionnaire explicatif et combinatoire du français contemporain – Recherches lexico-sémantiques III,* Montréal, Presses de l'Université de Montréal.

POLGUÈRE, A. (1990) : « Structuration et mise en jeu procédurale d'un modèle linguistique déclaratif dans un cadre de génération de texte », thèse, Département de linguistique et de philologie, Université de Montréal.

19

Connecteurs et traitement automatique

Gaston Gross

Université Paris XIII et INaLF, Paris, France

• *Abstract* •

This paper sets out how to process French connectors automatically. To do this, a new description of French conjonctive phrases is put forward. Instead of considering them as fixed sequences, the noun wich they contain is analysed as a predicate, with their main clause and the subordinate clause as arguments. The syntax of these noun predicates is described in detail.

Les mots de liaison

Il n'existe, à l'heure actuelle, que peu d'études d'ensemble sur ce qui, dans un texte donné, n'est ni opérateur ni arguments. Les études spécifiques sont nombreuses. Certains de ces travaux sont d'une grande précision comme celui de M. Piot sur les conjonctions ou ceux consacrés par les tenants de la pragmatique aux particules qui traduisent l'attitude du locuteur dans son propre discours. De même, les dictionnaires électroniques se sont surtout intéressés à la description des opérateurs dans le cadre de la phrase simple.

Peut-être cette absence d'études générales vient-elle de ce qu'il est difficile de donner un statut théorique à cet ensemble d'éléments hétéroclites. On parle à leur sujet de relateurs ou plus généralement d'adverbiaux. Mais cette démarche est exclusivement morphologique ou sémantique, peut-être parce qu'on est en présence d'objets qui échappent, pense-t-on, à une formulation sous forme de règles. Nous voudrions montrer dans cet article qu'une description précise de ces phénomènes est nécessaire si l'on veut comprendre leur fonctionnement et proposer des explications qui puissent faire l'objet, dans un deuxième temps, d'un traitement automatique.

La grammaire scolaire a résolu le problème de façon expéditive. Elle possède le concept de *composition.* Toutes les catégories grammaticales peuvent être simples ou composées. Elles ont, dans les deux cas, les mêmes propriétés syntaxiques. Ainsi *pomme de terre* et *patate, maître d'école* et *instituteur, joyeux* et *de bonne humeur, mourir* et *casser sa pipe, quand* et *au moment où* ont une syntaxe similaire et peuvent entrer dans un même paradigme. L'observation qui vient d'être faite à propos de la composition, si elle n'est qu'une approximation pour les catégories du nom, de l'adjectif et du verbe, constitue à coup sûr une erreur pour les conjonctions. Par exemple, on admettra volontiers que les mots de liaison des deux phrases suivantes sont en relation de quasi-synonymie :

Luc est arrivé quand nous sommes partis
Luc est arrivé au moment où nous sommes partis

Les restrictions sémantiques sont les mêmes pour *quand* et pour *au moment où.* Si l'on ajoute à ces connecteurs un adverbe comme *précisément* le résultat reste acceptable :

Luc est arrivé précisément quand nous sommes partis
Luc est arrivé précisément au moment où nous sommes partis

Mais les textes ne présentent pas toujours une correspondance aussi simple. Imaginons que l'adverbe *précisément* soit remplacé par l'adjectif *précis.* La disparité apparaîtra immédiatement :

*Luc est arrivé quand précis nous sommes partis
Luc est arrivé au moment précis où nous sommes partis

On voit donc que l'utilisation d'équivalents sémantiques peut, dans le cas des connecteurs, mener à des impasses et poser à la traduction automatique des difficultés majeures. Imaginons encore que, pour diverses raisons, nous ayons déjà parlé de l'heure de ce départ projeté. Cette information constitue donc pour notre interlocuteur un fait connu, par rapport auquel nous allons situer un autre événement, l'arrivée de Luc. Dans ce cas, nous pourrons dire très naturellement :

Luc est arrivé à ce moment

où le démonstratif *ce* réfère au départ en question. Cette expression ne peut pas non plus avoir de correspondant avec la forme en *quand,* qui n'est pas un substantif.

On voit donc qu'il est impossible d'assimiler les deux relateurs bien que, dans certaines conditions, ils puissent être synonymes. Pour pouvoir traduire ou plus généralement traiter automatiquement toutes les suites comprenant des connecteurs, il faut à la fois une théorie des connecteurs et des descriptions plus poussées. Cet article est consacré à une classe particulière de connecteurs : les locutions conjonctives.

Description des locutions conjonctives

Parmi les locutions conjonctives nous portons notre attention à celles qui com-

prennent un substantif. Nous allons montrer que ces locutions ont un fonctionnement beaucoup plus complexe que les équivalences pédagogiques ne le laissent entendre. L'analyse traditionnelle, peut-être à cause d'une analogie avec les noms composés – considérés comme des unités au même degré que les noms simples, en raison de la non-compositionnalité de leur sens – assimile, dans l'analyse, conjonctions et locutions conjonctives.

Au lieu de considérer les locutions comme des ensembles figés qui ne relèvent pas d'une analyse syntaxique, nous portons notre attention sur le substantif que nous considérons comme l'élément essentiel de la locution. Les autres constituants sont respectivement la préposition et la détermination de ce substantif. Cette détermination comprend un prédéterminant et un modifieur. Nous sommes donc en présence des quatre éléments suivants :

Prép Dét N Modif

Cette structure n'est pas figée, comme le laisse entendre l'analyse habituelle qui assimile dans les faits, comme nous l'avons dit, les locutions conjonctives aux conjonctions simples. Nous allons montrer que tous ces éléments peuvent constituer des paradigmes. Cette possibilité constitue un critère de non-figement.

La préposition

La préposition peut faire l'objet d'un choix. On observe ainsi avec certains substantifs les substitutions suivantes :

à/avec
    à l'aide de/avec l'aide de
avec/dans
    avec le but de/dans le but de
    avec l'intention de/dans l'intention de
de/en
    de sorte que/en sorte que
par/de
    par crainte de/ de crainte de
par/zéro
    par crainte de VW/crainte de VW

Il arrive que la préposition puisse être effacée dans certaines conditions, en particulier quand le modifieur est une phrase :

Luc s'est tu (par + E) crainte de se faire gronder

Le substantif

Dans beaucoup de cas, le substantif peut former un paradigme sémantiquement homogène :

avec LE (souci, volonté, désir, espoir,...) de VW
à LE (moment, instant, seconde, minute, heure) où P
pour LE (motif, raison) que P
de (sorte, manière, façon) que P

Il est aisé de multiplier les exemples. Nous verrons plus loin que le substantif détermine, en fait, les changements possibles de la préposition et de la détermination.

Les déterminants

Dans la mesure où la détermination du substantif qui figure dans les locutions conjonctives est constituée d'une détermination composée spécifique notée *Dét-Modif* (c'est-à-dire un prédéterminant suivi d'un modifieur correspondant à la circonstancielle), il est impossible d'étudier séparément les deux éléments de cet ensemble. Nous analyserons donc en détail cette détermination. On verra qu'au terme de cette analyse le caractère schématique de l'analyse scolaire traditionnelle apparaîtra clairement.

Types de déterminants

La détermination affirmative

Nous regroupons sous ce terme l'ensemble des déterminants affirmatifs. Ce type de détermination est constitué de deux ensembles formés par la cataphore et l'anaphore.

Détermination cataphorique

La cataphore comprend des unités qui réfèrent à des éléments à venir, contrairement à l'anaphore qui renvoie à des éléments figurant dans un contexte de gauche. À titre d'exemple, l'article *le* est dit anaphorique quand il réfère à un objet déjà identifié par le contexte ou la situation, comme dans la phrase :

Redonne-moi le livre

Ici le substantif *livre* est identifié par la situation que crée le préfixe *re-* référant à un événement qui s'est produit une première fois. Hors d'une situation semblable, le substantif ne peut pas être identifié. Si donc on n'a jamais parlé de livre, comme dans *Donne-moi le livre* prononcé hors contexte, il est impossible qu'un interlocuteur puisse prendre cette phrase comme une injonction qui lui est adressée, alors que c'est possible, dans les mêmes conditions, pour la suivante :

Donne-moi le livre qui est sur la table

Dans cette dernière phrase l'article *le* ne renvoie pas à ce qui précède mais « annonce » la relative *qui est sur la table.* Il constitue donc un emploi cataphorique. À la place de *le*, on peut noter aussi l'article indéfini et l'article zéro :

Le-Modif
Luc a dit cela dans le but de convaincre
Luc a fait cela dans l'espoir qu'on le comprendra
Luc est resté du fait qu'il pleuvait
Luc s'est tu pour la raison qu'il ignorait tout

Un-Modif
Luc a dit cela dans un souci d'apaisement
Luc a dit cela dans un but évident de convaincre

E-Modif
Luc a dit cela sous prétexte d'informer son voisin
Luc a dit cela afin de plaire à son père

On voit que la détermination cataphorique représente en fait les constructions appelées *propositions subordonnées circonstancielles.* Notons tout de suite que la subordonnée circonstancielle peut revêtir trois formes syntaxiques : la forme conjuguée,

afin que cela soit expertisé

la réduction infinitive,

afin d'expertiser cela

la nominalisation,

à des fins d'expertise

Nous appelons ces modifieurs *modifieurs complétifs* parce qu'ils sont formés d'une phrase. Il faut ajouter qu'à côté de cette détermination complétive, il existe des modifieurs de type adjectival. On peut avoir, par conséquent, deux niveaux de modifieurs comme on le voit dans l'exemple suivant :

Luc est parti pour la raison qu'il était malade
Luc est parti pour la raison évidente qu'il était malade

Nous obtenons, en faisant une étude systématique, les modifications suivantes portant sur la détermination. Nous notons *Mc* le modifieur complétif (en fait la circonstancielle) et *Ma* le modifieur adjectival.

E-Mc/le-Mc
sous prétexte que P
sous le prétexte que P

E-Mc/E-Ma-Mc
afin que P
à seule fin que P

en fonction de N
en fonction inverse de N

le-Mc/ce-Mc
pour la raison que P
pour cette raison que P

le-Mc/le-Ma-Mc
dans le but de VW
dans le seul but de VW

le-Mc/un-Ma-Mc
dans le but de VW
dans un but évident de VW

le-Mc/le-Ma : P
dans le but de rendre justice à Luc
dans le but que voici : rendre justice à Luc

pour la raison qu'il pleuvait trop fort
pour la raison suivante : il pleuvait trop fort

Détermination anaphorique

Nous venons de voir que, dans le cas des subordonnées circonstancielles, la détermination du substantif est cataphorique : elle annonce les « circonstances » représentées par ces subordonnées. Il se trouve, comme nous l'avons vu plus haut, que ces « circonstances » peuvent être déjà connues de l'interlocuteur. Dans ce cas, on réfère à elles par des déterminants *anaphoriques.* Voici l'essentiel de ces déterminants.

Pronominalisation du modifieur

Le modifieur complétif, c'est-à-dire la subordonnée circonstancielle, peut faire l'objet d'une pronominalisation dont le but est de référer à un fait antérieur connu. Ce pronom a différentes formes : le *relatif de liaison,* les pronoms *cela, ça* et *là*

le relatif de liaison :
en conséquence de ce que P/en conséquence de quoi
au lieu de ce que P/au lieu de quoi
faute de ce que P/faute de quoi
après que P/après quoi

les pronoms *cela* et *ça*
en comparaison de ce que P/en comparaison de (cela + ça)
en conséquence de ce que P/en conséquence de (cela + ça)
après que P/après (cela + ça)
à cause de ce que P/à cause de cela

le pronom *là*
d'ici à ce que P/d'ici là
en fonction de ce que P/en fonction de là

Il semble que, dans la cas de l'anaphore, on puisse aussi effacer le modifieur :

En attendant (E + qu'il vienne), je lirai
À défaut (E + de pouvoir lire), je me promènerai
À force (E + de crier comme ça), tu te fatigueras

Le démonstratif

Le déterminant cataphorique *le-Mc* peut être remplacé par le démonstratif *ce*, qui renvoie à un événement antérieur considéré comme connu. Il s'agit du cas d'anaphore le plus fréquent. En voici quelques exemples :

À cette fin, il faut s'inscrire à la mairie
Dans ces circonstances, je ne partirai pas
Dans ce but, tu consulteras un notaire
Dans ces conditions, il est impossible de travailler
De ce fait, le travail n'a pas pu être fait
Pour cette raison, les travaux se sont arrêtés
Dans ce cas, il est inutile de continuer

On observera cependant une différence de comportement syntaxique entre les déterminants cataphoriques et les anaphoriques. Alors que les premiers réfèrent à des phrases seulement (les subordonnées), les secondes peuvent référer aussi à un contexte ou à une situation.

Adjectifs

Certains adjectifs permettent, de la même façon que le démonstratif, de référer à un événement antérieur connu, avec quelquefois une interprétation plus générique :

dans un tel but
dans des circonstances pareilles
dans un but de ce genre
dans pareille circonstance
pour de pareilles raisons
pour des raisons similaires
pour des raisons de ce genre
pour des raisons semblables

L'adjectif *précédent* semble être ici un prototype, tout comme *suivant* l'est pour la cataphore.

Autres modifieurs

La référence à un événement connu de l'interlocuteur peut se faire à l'aide d'adjectifs spécifiques ou de propositions relatives.

pour la raison susmentionnée
pour le motif ci-dessus

pour la raison que je viens de dire
pour la raison qui vous est connue
pour la raison que vous avez lue

Déterminants négatifs

Ce qui montre avec plus de clarté encore que les structures que nous étudions ne sont pas figées, c'est la possibilité de mettre devant ces substantifs des déterminants négatifs :

pour la raison que/pour aucune autre raison que
au moment où P/à aucun moment
dans le but précis de VW/dans aucun but précis
pour le motif que P/pour nul autre motif

Déterminants interrogatifs et exclamatifs

pour quelle raison ?
pour quel motif ?
à quel moment ?
dans quelles conditions ?
dans quel but ?

et dans quelles conditions !
et à quelle fin !
et à quel moment !

Déterminants « indéfinis »

à certaines fins indéterminées
à toutes fins utiles
pour d'autres raisons
à un certain moment
dans d'autres circonstances
pour n'importe quelle raison
pour telle ou telle raison

Déterminants quantifieurs

Un certain nombre de « locutions » acceptent un déterminant quantifieur devant le substantif :

pour deux raisons
pour plusieurs raisons : la première... la seconde...
pour les deux motifs suivants

La liberté de détermination des substantifs que nous venons de noter montre clairement que nous n'avons pas affaire à des conjonctions composées, au sens où l'on parle de noms composés, puisque le substantif qui y figure a une syntaxe spécifique que nous allons maintenant mettre en évidence.

Les substantifs comme opérateurs

Nous venons de montrer, dans ce qui précède, que les locutions conjonctives ou prépositives ne peuvent pas être assimilées à des catégories grammaticales comme les conjonctions simples *si, quand, lorsque,* etc. Nous avons vu aussi que le substantif qui y figure est le terme le plus important et qu'il est accompagné de deux autres éléments qui dépendent de lui : la détermination et une préposition. Nous allons décrire la syntaxe de ces différents éléments.

Statut du substantif

Observons d'abord que ces substantifs sont des noms abstraits. Les travaux du LADL ont montré que les abstraits sont des prédicats nominaux. Comme tels, ils ont des arguments. Nous verrons ce fait dans un instant. Notons que les substantifs prédicatifs sont actualisés par un verbe support. Ce verbe support dépend de la nature du substantif prédicatif. Avec un nom comme *voyage,* on aura les supports *faire* (ou l'une de ses variantes comme *effectuer, réaliser,* etc.) ou *être en* :

Luc fait un voyage en Auvergne
Luc est en voyage en Auvergne

Les substantifs *envie, réponse, opposition, conclusion* ont, par exemple, les supports suivants :

Luc a envie de partir
Luc a donné une réponse
Luc est en opposition avec Pierre
Luc a tiré la conclusion que tout était faux

Les éléments nominaux que nous observons dans les locutions conjonctives sont donc actualisés par des verbes supports. Les substantifs qui traduisent un « but » (*but, intention, désir, espoir,* etc.) sont accompagnés du verbe support *avoir* :

avec le but de/avoir le but de
avec l'intention de/avoir l'intention de
avec le désir de/avoir le désir de
avec l'espoir de/avoir l'espoir de

Dans le cas des subordonnées finales, il est clair que ces prédicats nominaux ont comme sujets le sujet de la « principale » :

Luc a dit cela, il avait l'intention de se disculper
Luc a dit cela avec l'intention de se disculper

Luc a dit cela, il avait le désir de se disculper
Luc a dit cela avec le désir de se disculper

Nous tirons donc une première conclusion : les locutions conjonctives finales reposent sur des phrases qui ont les caractéristiques suivantes :

- le prédicat de cette phrase est constitué d'un substantif ;
- le sujet de ce prédicat est coréférent à celui de la phrase dite principale ;
- ce substantif est actualisé par un verbe support (*avoir*) qui est spécifique de cette classe de prédicats nominaux ;
- ce verbe support peut avoir une variante non actualisée représentée par une préposition : *avec* est relié au support *avoir* ;
- les structures finales reposent donc, en définitive, non sur deux phrases, la principale et la subordonnée finale, mais sur trois phrases, la principale, la « subordonnée » et celle qui est cachée derrière la locution conjonctive et dont l'opérateur est un prédicat nominal traduisant « l'intention » du sujet de la principale.

Cette analyse permet seule de relier entre elles toutes les possibilités d'expression qu'un substantif prédicatif donné permet de générer pour exprimer le but. Prenons la racine prédicative *désir-*. On sait que ce prédicat peut avoir une forme nominale, verbale ou adjectivale. On aura donc les formes suivantes :

Luc désire s'expliquer
Luc a le désir de s'expliquer
Luc est désireux de s'expliquer

Chacune de ces formes peut faire l'objet d'un certain nombre de modifications morpho-syntaxiques. Pour ce qui est de la forme verbale, on peut avoir une relativation ou une forme participiale :

Luc, qui désire s'expliquer
Luc, désirant s'expliquer

La forme nominale est caractérisée par le support *avoir* et ses variantes :

Luc a le désir de s'expliquer

Luc caresse le désir de s'expliquer
Luc nourrit le désir de s'expliquer

La relativation peut s'appliquer à ces phrases :

le désir que Luc a de s'expliquer

Remarquons que le groupe nominal ainsi formé comporte encore un prédicat nominal actualisé grâce au support *avoir* qui figure dans la relative. Cette actualisation peut être supprimée par l'effacement du support et par l'effacement concomitant du relateur *que* :

Le désir de Luc de s'expliquer

Le complément de nom est susceptible de pronominalisation à l'aide de l'adjectif possessif :

son désir de s'expliquer

Une autre façon d'enlever l'actualisation consiste à remplacer le verbe *avoir* par les prépositions *avec* ou *dans* :

avec le désir de s'expliquer
dans son désir de s'expliquer

La forme adjectivale peut donner lieu aux transformations suivantes : formation d'apposition par effacement du substantif *être* ou adjonction d'un adverbe intensif :

Luc, désireux qu'il est de s'expliquer
Luc, désireux de s'expliquer
Luc, très désireux de s'expliquer

Quelle que soit la forme de la racine *désir-*, le mot construit sur cette racine sera un opérateur. Cet opérateur a comme argument-sujet un nom humain et comme argument-objet une phrase. Cette phrase représente la motivation qui pousse un sujet à faire une action déterminée. On peut donc postuler comme point de départ la structure suivante :

Luc s'est levé. Il désirait s'expliquer

avec ces modifications :

Luc, qui désirait s'expliquer, s'est levé
Luc, désirant s'expliquer, s'est levé

Si l'on adopte la forme adjectivale, on aura alors :

Luc s'est levé. Il était désireux de s'expliquer
Luc, désireux qu'il était de s'expliquer, s'est levé
Luc, désireux de s'expliquer, s'est levé

La forme nominale est caractérisée par le verbe support *avoir* :

Luc s'est levé. Il avait le désir de s'expliquer

Les transformations suivantes s'appliquent à ces phrases :

Luc s'est levé, avec le désir de s'expliquer
Luc s'est levé, dans son désir de s'expliquer

En résumé, nous considérons que des substantifs comme *désir* sont des opérateurs. Cette analyse est confortée par les autres positions syntaxiques que peut avoir le substantif désir.

Le désir de Luc en se levant était de s'expliquer
S'expliquer était le désir de Luc
Luc avait comme désir de s'expliquer
Ceci était le désir de Luc
Ceci était son désir

Pour pouvoir traiter de façon automatique les constructions finales de ce type, il faut être en mesure de rendre compte de toutes les relations qui relient les constructions de l'opérateur dont le radical est *désir-*. Ce traitement n'est possible qu'à certaines conditions théoriques. Il faut pour cela avoir dans son appareillage théorique les éléments suivants :

- il y a des substantifs prédicatifs ;
- ces substantifs prédicatifs sont actualisés par des verbes supports ;
- les verbes supports peuvent avoir certaines variantes « prépositionnelles » comme *avec* par rapport à *avoir* ;
- ces substantifs prédicatifs ont deux arguments, représentés respectivement par la principale et la « subordonnée ».

On voit donc que les locutions conjonctives ne peuvent pas être assimilées, en profondeur, à une catégorie grammaticale comme la conjonction. La tâche du linguiste est de mettre au point toutes les propriétés syntaxiques des opérateurs comme *désir-*. On verra alors, en observant les paraphrases, que ce qu'on appelle *locution conjonctive* n'est autre que l'une des étapes transformationnelles qui caractérisent cet opérateur.

Autre exemple d'analyse : le substantif « cause »

Nous partons de la construction verbale mettant en jeu le verbe *causer.* Pour des problèmes de lisibilité, nous n'employons pas d'arguments propositionnels car cela alourdirait inutilement les phrases, mais nous les remplaçons par les pronoms phrastiques *ceci* et *cela.* Nous partons de la phrase verbale :

Ceci a causé cela

Nous appliquons à cette phrase la transformation passive :

Cela a été causé par ceci

Il existe des verbes supports qui nominalisent des passifs verbaux. C'est le cas, entre autres, des supports *recevoir, subir, faire l'objet de* :

Luc a été admonesté par Paul
Luc a reçu des admonestations de la part de Paul

Luc a été brimé par Paul
Luc a subi des brimades de la part de Paul

On a procédé à un remaniement de ce service
Ce service a subi un remaniement

Parmi ces support passifs on peut classer *être à* que l'on trouve dans :

Ce travail est programmé
Ce travail est au programme

Nous appliquons ce support au passif que nous avons obtenu :

Cela a été causé par ceci
Cela a été à cause de ceci

Si nous effaçons le support, ce qui est une propriété générale de ces verbes, on obtient :

Cela, à cause de ceci

Nous avons ainsi expliqué la source de la locution prépositive *à cause de,* ou plutôt, nous avons mis en relation les différents emplois de la racine *caus-*.

Le problème du figement

Nous avons étudié jusqu'à présent les locutions conjonctives en prenant le contre-pied de la position habituelle. Nous avons, dans un premier temps, postulé que nous sommes en présence de structures nominales qui ont des régularités syntaxiques. Ceci ne veut pas dire cependant que toutes les locutions conjonctives aient toutes l'ensemble des propriétés que nous venons de décrire. La syntaxe des substantifs prédicatifs est plus ou moins libre. Elle est soumise à des restrictions et à des degrés divers de figement. Le travail de description qu'il convient de faire pour le traitement automatique consiste, pour chaque substantif considéré comme prédicat de deuxième niveau, à décrire les impossibilités par rapport aux régularités que nous avons indiquées.

Prédicats nominaux seulement

Nous venons de voir l'exemple de la racine *désir-* qui a les formes nominale, verbale et adjectivale et celui de *cause,* qui est à la fois verbe et nom mais n'a pas de forme adjectivale. Des substantifs comme *volonté* n'ont que la forme verbale et nominale :

Il veut s'expliquer
Il a la volonté de s'expliquer
*Il est volontaire de s'expliquer

quant à *intention,* il n'a que la forme nominale, à l'exclusion de la forme verbale et adjectivale :

Il a l'intention de s'expliquer
*Il intentionne de s'expliquer
*Il est intentionné de s'expliquer

On ne trouve que la forme nominale aussi avec *dessein* :

Il a le dessein de s'expliquer

Dans ce cas, les possibilités de paraphrases, c'est-à-dire la diversité des moyens d'expression de la langue, sont plus réduites. Ces lacunes lexicales sont, en français et

en anglais par exemple, beaucoup plus grandes que dans des langues sémitiques comme l'arabe ou l'hébreu. Cependant, on peut créer des néologismes en forçant l'acceptabilité d'une forme morphologique. Ainsi, on serait en mesure de comprendre les phrases suivantes et de les considérer comme non acceptables mais non totalement agrammaticales :

Il intentionne de s'expliquer

Restrictions portant sur la transformation du modifieur

Le modifieur, qui est de nature phrastique, peut revêtir trois formes. Il peut être constitué d'une phrase conjuguée (la circonstancielle), d'une infinitive (la réduction infinitive de cette circonstancielle) ou d'une nominalisation. Ainsi le substantif prédicatif *fin* peut avoir ces trois formes, dans certaines conditions, tout comme le substantif *fait* :

On a emporté ce tableau afin qu'il soit expertisé
On a emporté ce tableau afin de l'expertiser
On a emporté ce tableau à des fins d'expertise
Luc est revenu du fait qu'il s'est trompé
Luc est revenu du fait de s'être trompé
Luc est revenu du fait d'une erreur

On observe toutes les restrictions potentielles. Ainsi la circonstancielle en *que* peut être interdite. C'est le cas avec le substantif *but* :

?Luc a dit cela dans le but que tous comprennent
Luc a dit cela dans le but d'expliquer la situation
Luc a dit cela dans un but d'explication

Le substantif *cause* admet un modifieur nominal mais ni l'infinitif ni, selon les puristes, un modifieur « phrastique » :

Luc est rentré à cause de la pluie
*Luc est rentré à cause de pleuvoir
?Luc est rentré à cause qu'il pleut

Le substantif *raison* a un comportement différent selon la préposition qui l'introduit :

Luc s'est repris pour la raison qu'il s'est trompé
*Luc s'est repris pour la raison de s'être trompé
?Luc s'est repris pour la raison de son erreur
Luc s'est égaré en raison d'une erreur
*Luc s'est égaré en raison qu'il a fait une erreur
*Luc s'est égaré en raison d'avoir fait une erreur

Rappelons pour mémoire que le modifieur peut aussi être effacé dans certaines conditions :

Après (E + qu'il eut dit cela) il est parti

Restrictions portant sur la détermination anaphorique

Nous avons vu plus haut que la détermination anaphorique est possible si l'indication de la « circonstance », au lieu de suivre comme dans le cas des subordonnées circonstancielles, est déjà connue, soit qu'elle fasse l'objet d'informations relevant de la situation, soit qu'elle réfère à une phrase précédente.

Impossibilité du démonstratif

Un très grand nombre de substantifs prédicatifs jouant le rôle de « connecteurs » peuvent avoir une détermination démonstrative :

afin que P/à cette fin
dans le but de VW/dans ce but
dans des conditions où P/dans ces conditions
du fait que P/de ce fait
pour la raison que P/pour cette raison
au moment où P/à ce moment

Cette détermination n'est pas possible dans un certain nombre de cas :

en raison de la pluie/*en cette raison
en dépit de Luc/*en ce dépit
en comparaison de cela/*en cette comparaison

Exemple de description d'un opérateur de deuxième niveau : le substantif but (avec le but de)

Changement de préposition

Dans cette position on peut trouver deux prépositions *avec* et *dans*

avec le but de
dans le but de

Détermination de ce substantif

Détermination cataphorique

Pour rendre lisible le tableau qui suit, nous devons donner quelques explications. Nous avons parlé, du point de vue de la détermination, de *modifieurs* et nous avons désigné par ce terme la phrase entière ou transformée qui est appelée proposition subordonnée circonstancielle. Nous appelons désormais ce modifieur un *modifieur complétif (Mc)* car il se trouve qu'on peut adjoindre un second type de modifieurs de type adjectival que nous appelons *modifieur adjectival (Ma)*. Ainsi dans la suite :

dans le but évident de réussir

le modifieur complétif est *de réussir* et le modifieur adjectival est *évident.* On a donc :

Mc = modifieur complétif (que P, de VW, de N)
Ma = modifieur adjectival

Notons encore que le symbole *E* désigne le déterminant zéro. On obtient ainsi pour la détermination cataphorique :

	que P	de VW	de N
LE-Mc	–1	+2	–3
Le-Ma-Mc	–4	+5	–6
Un-Mc	–7	–8	+9
Un-Ma-Mc	–10	–11	+12
E-Mc	–13	–14	–15
E-Ma-Mc	–16	–17	–18

(1) *dans le but qu'il réussisse
(2) dans le but de réussir
(3) *dans le but de réussite
(4) *dans le but avoué qu'il réussisse
(5) dans le but avoué de réussir
(6) *dans le but avoué de réussite

(7) *dans un but qu'il réussisse
(8) *dans un but de réussir
(9) dans un but de réussite
(10) *dans un but avoué qu'il réussisse
(11) *dans un but avoué de réussir
(12) dans un but avoué de réussite

(13) *dans but qu'il réussisse
(14) *dans but de réussir
(15) *dans but de réussite
(16) *dans but avoué qu'il réussisse
(17) *dans but avoué de réussir
(18) *dans but avoué de réussite

Détermination anaphorique

Si le but poursuivi par quelqu'un a déjà fait l'objet d'une discussion ou si la situation a apporté cette information alors on peut avoir comme déterminant du substantif *but* les éléments suivants :

a) le démonstratif : *dans ce but*
b) certains adjectifs : *dans un tel but, dans un pareil but, dans pareil but, dans un but semblable*

Autres déterminants

Comme nous l'avons signalé plus haut, on peut trouver des déterminants,

a) négatifs : *dans aucun but (précis, particulier)*
b) interrogatifs : *dans quel but ?*
c) exclamatifs : *et dans quel but !*

Utilisations de verbes supports

Comme nous l'avons vu, le substantif *but* est un prédicat nominal. En tant que tel, il peut être actualisé par un verbe support, en l'occurrence le support *avoir* :

Luc a fait cela avec le but de réussir
Luc a fait cela. Il avait le but de réussir

Il se pourrait que ce verbe ait une variante en *avoir comme* :

Luc a comme but de réussir

On trouve un autre verbe sur le statut duquel nous ne voudrions pas nous prononcer ici. Il s'agit du verbe approprié *se fixer* :

Luc s'était fixé le but de réussir
Luc s'était fixé comme but de réussir

Il semble qu'il soit à rapprocher d'un verbe comme *donner* :

Luc s'était donné le but de réussir
Luc s'était donné comme but de réussir

En position prédicative

L'analyse que nous proposons des substantifs qui figurent dans les locutions conjonctives comme étant des prédicats est confortée par la possibilité qu'ont ces substantifs de fonctionner en position prédicative. Ce phénomène est observable dans des phrases comme :

Réussir était le but de Luc
Luc avait comme but de réussir
Ceci constituait un but pour Luc

Syntaxe du substantif fin de la locution « afin que »

Nous décrivons, dans ce qui suit, le comportement syntaxique du substantif prédicatif *fin*. On notera que par rapport au substantif *but* il a un comportement plus restreint.

Changement de la préposition

Le préposition *à* qui figure dans *afin que* ne peut pas être remplacée.

Détermination du substantif

Détermination cataphorique

	que P	de VW	de N
Le-Mc	–1	–2	+3
Le-Ma-Mc	+4	+5	–6
Un-Mc	–7	–8	+9
Un-Ma-Mc	–10	–11	+12
E-Mc	+13	+14	–15
E-Ma-Mc	+16	+17	–18

(1) *à la fin qu'il réussisse
(2) *à la fin de réussir
(3) *à la fin de réussite mais aux fins de réussite
(4) à l'unique fin qu'il réussisse
(5) à l'unique fin de réussir
(6) *à l'unique fin de réussite

(7) *à une fin qu'il réussisse
(8) *à une fin de réussir
(9) à (une + des) fin(s) de réussite
(10) *à une fin avouée qu'il réussisse
(11) *à une fin avouée de réussir
(12) à une fin avouée de réussite

(13) afin qu'il réussisse
(14) afin de réussir
(15) *afin de réussite
(16) à seule fin qu'il réussisse
(17) à seule fin de réussir
(18) *à seule fin de réussite

Détermination anaphorique

a) démonstratif

afin que P/à cette fin

b) adjectifs

afin que P/à (telle + pareille) fin

Autres déterminants

a) déterminants négatifs

à aucune fin précise
à nulle autre fin

b) déterminants interrogatifs

à quelle fin ?

c) déterminants exclamatifs

et à quelle fin !

Utilisation de verbes supports

En français moderne, il n'y a pas de vrais supports avec le substantif prédicatif *fin.* Dans la langue classique, le support était naturellement *avoir.*

Autres verbes

Luc s'est fixé pour fin de réussir
Luc s'est proposé comme fin de réussir

Constructions en position prédicative

Réussir était une fin pour Luc
Réussir était la fin qu'il s'était proposée

Les travaux que nous avons entrepris consistent à décrire avec la même précision tous les substantifs qui constituent des opérateurs de deuxième niveau. Il s'agit de plus de mille prédicats nominaux.

20

Identification et codage des phraséologismes verbaux dans un environnement de traduction automatique

Marie-Claude L'Homme

Lexi-tech Inc., Moncton (N.-B), Canada

• *Abstract* •

Over the past ten years, phraseology has gained an increasing interest in terminology. Although existing models and proposals are still in their infancy, studies in phraseology offer a wide variety of applications. One of them is machine translation.

The linguistic context has rarely been taken into account in the description of terms. However, automatic processing of terminological units illustrates the limits of traditional models based exclusively on the conceptual representation of terms.

This paper deals with one aspect of phraseology, that is the identification and representation of verbal combinations in technical language. It will demonstrate how the coding of syntactic patterns can solve ambiguity problems in machine translation environment.

Introduction

Le thème de la phraséologie terminologique suscite, chez les spécialistes et praticiens en quête de solutions nouvelles, un intérêt grandissant. Après avoir concentré leurs efforts de systématisation sur les rapports entre notion et forme, les spécialistes de la terminologie s'attardent maintenant à la réalisation linguistique des termes en contexte.

Ce nouveau champ d'étude remet en question l'application exclusive des modèles théoriques rigides sur lesquels reposent les principes de confection d'outils ter-

minographiques (lexiques scientifiques et techniques élaborés selon les règles de l'art et banques de terminologie). Il tient compte d'éléments d'analyse et de facteurs qui n'étaient pas considérés auparavant, comme les unités lexicales autres que nominales, le comportement des termes en contexte, etc.

Bien qu'elle en soit encore à ses balbutiements et qu'elle n'ait qu'un statut théorique incertain, l'étude du contexte linguistique spécialisé présente des applications pratiques incontestables. La traduction automatique (TA) constitue l'une de ces applications. Le traitement automatique de l'unité terminologique illustre parfaitement, non seulement l'intérêt de ce genre d'étude, mais sa nécessité. La manipulation de l'unité terminologique par un système de TA démontre les limites des modèles traditionnels fondés sur le principe de l'univocité. Par ailleurs, bien que les recherches actuelles insistent fréquemment sur une représentation sémantique adéquate du lexique dans les bases de données lexicales, il s'avère que le traitement des mots et des termes ne peut échapper à une étude de leur comportement en contexte.

Le présent article s'attardera sur l'un des aspects de la phraséologie spécialisée, à savoir les différentes méthodes de codage des phraséologismes verbaux de la langue technique. Dans un contexte plus général, il illustrera de quelle façon une analyse des liens syntaxiques entre les mots peut résoudre certains problèmes d'ambiguïté entraînés par l'association exclusive de formes linguistiques dans une base de données lexicales[1].

Traduction automatique et terminologie

L'unité lexicale univoque est certainement celle qui se prête le mieux à la traduction automatique telle qu'on la connaît aujourd'hui. Par exemple, un terme comme *cathode ray tube,* quel que soit le texte, quelle que soit la phrase dans laquelle il apparaît ou sa place dans cette phrase, se traduira toujours par *tube à rayons cathodiques* (ou l'équivalent préconisé par le traducteur).

Ces circonstances idéales sont valables pour certaines catégories d'unités lexicales, c'est-à-dire les termes nominaux et adjectivaux très spécialisés (*alphagraphic - alphagraphique ; optical isolator - photocoupleur*) ou encore, certains mots non ambigus appartenant à la langue technique (*wattmeter - wattmètre ; hertz - hertz*). En fait, il s'agit de mots qui, quel que soit le contexte, se traduiront toujours de la même façon.

Dans ces cas précis, il s'agit simplement de coder les différentes formes et de leur assigner un équivalent dans la base de données lexicales d'un système de TA. Celle-ci se présente comme un inventaire plus ou moins complet des formes que le logiciel peut rencontrer dans un texte de départ, formes auxquelles sont associées des unités à reproduire dans un texte d'arrivée. Cela suppose donc que plus la base est complète (et *complet* n'implique par uniquement la notion de *volumineux*), plus la qualité de la traduction machine sera satisfaisante.

Les quelques exemples cités ci-dessous (tableau 1) permettront de décrire, de façon très succincte, le fonctionnement de la base de données lexicales.

1. Les différents problèmes et exemples traités dans le présent article s'appuient sur des expériences concrètes menées dans une entreprise qui se spécialise dans la traduction de textes techniques. *Lexi-Tech* (nom de l'entreprise) exige de la part de tous ses traducteurs qu'ils postéditent une sortie-machine.

TEXTE DE DÉPART	BASE DE DONNÉES	TEXTE D'ARRIVÉE
formes de départ	forme anglaise - forme française	formes d'arrivée
blind nut, blind nuts	blind nut - écrou borgne (nom masculin variable)	écrou borgne, écrous borgnes
asynchronous	XXXX	???
bus bar	1. bus - bus (nom masculin invariable) ; 2. bar - barre (nom féminin variable)	barre de bus (*barre omnibus)

Tableau 1.

Une forme contenue dans un texte de départ peut apparaître (*blind nut,* par exemple) ou non (*asynchronous*) dans la base de données lexicales. Dans le premier cas, elle est reproduite dans le texte d'arrivée (*écrou borgne*) ; dans le second, elle n'est simplement pas reconnue par le système de TA (*???*). S'il s'agit d'un terme complexe (*bus bar*), l'expression comme telle peut ne pas apparaître dans la base de données mais chacune de ses composantes (*bus* et *bar*) peuvent y figurer. Elles seront donc traduites séparément (*barre* et *bus*) et liées par le connecteur implicite (*barre de bus*) jusqu'à ce que le terme entier ait été ajouté à la base (*barre omnibus).*

Le texte à haute densité terminologique, ou plutôt à grande concentration d'unités monosémiques, est certes celui que la TA traite le mieux. De plus, le système de TA assure une uniformité à laquelle la traduction plus traditionnelle ne peut aspirer. Une fois qu'un terme a été ajouté à la base de données lexicales, non seulement le logiciel de traduction reconnaît la forme dans la langue source et donne l'assurance de reproduire l'équivalent correct, mais il le fera, avec une régularité presque assommante, il convient de le dire, d'un bout à l'autre du texte, d'un texte à l'autre pour un même utilisateur et d'un utilisateur à l'autre.

Traduction automatique et polysémie

Cependant, certains problèmes surviennent car tout texte, et même le texte technique, mais dans une moindre mesure, comporte des unités polysémiques. Les problèmes les plus fréquents sont les suivants :

1. unités lexicales ayant un sens terminologique et un sens général qui peuvent tous les deux se retrouver dans le texte technique (par exemple, *thread - fil, filet ; leak - fuite, voie d'eau)* ;
2. unités terminologiques ayant des sens distincts dans deux ou trois domaines différents (par exemple, *operator* qui peut être un *téléphoniste* en télécommunications ou un *opérateur* en informatique ; *overflow* qui renvoie à *dépassement* en informatique et à *débordement* en télécommunications) ;
3. unités terminologiques ayant des sens distincts dans un même domaine (*control valve : robinet de réglage* ou *vanne de régulation* en robinetterie) ;
4. mots de la langue technique qui ne peuvent être associés à un domaine de spécialité et qui sont polysémiques (*console* qui peut être un *pupitre* ou une *console*

selon la taille de l'objet ; *clear* qui peut se rendre par *remettre à zéro*, *supprimer*, etc. ; *support*, par *prendre en charge, être conforme à*, etc.)[2].

Chacun des exemples qui précèdent fait ressortir un problème linguistique particulier mais tous présentent un point commun sur le plan de la traduction. Il faut, pour interpréter le sens de ces unités, considérer le contexte dans lequel elles apparaissent.

Contexte et traduction automatique

L'analyse du contexte, pour le traducteur, signifie la prise en compte d'éléments aussi divers que l'environnement linguistique immédiat et plus vaste dans lequel figure l'unité ou le contexte général dans lequel se situe le texte ; la reconnaissance de distinctions sémantiques en s'appuyant sur des outils de travail (livres) ou ses propres connaissances, etc. Dans certains cas, cette analyse se fait machinalement chez le traducteur d'expérience.

Pour un logiciel de TA, qui procède toujours d'une forme linguistique, il existe, dans l'état actuel des connaissances, deux façons d'établir des distinctions sémantiques.

La première méthode consiste à discriminer les sens d'une même forme linguistique en associant chacune de ses significations à un domaine de spécialité. En fait, il s'agit d'établir une association exclusive entre deux formes linguistiques en limitant cette association à un domaine de connaissances.

Par exemple (tableau 2), le mot *connector* peut être codé une première fois dans le domaine de la mécanique avec l'équivalent *raccord* et une seconde fois dans le domaine de l'électrotechnique avec l'équivalent *connecteur*. Il s'agira de préciser au logiciel de TA le domaine approprié lors du traitement du texte.

TEXTE DE DÉPART	BASE DE DONNÉES LEXICALES	DOMAINE SÉLECTIONNÉ	TEXTE D'ARRIVÉE
1. (domaine traité : mécanique) The connector is located under the main unit.	1. (mécanique) raccord, nom masculin variable	1. Mécanique	1. Le raccord se trouve sous l'unité principale.
2. (domaine traité : électrotechnique) The connector is located under the main unit.	2. (électrotechnique) connector, nom masculin variable	2. Électrotechnique	2. Le connecteur se trouve sous l'unité principale.

Tableau 2.

2. Cette catégorie de termes pose les plus grands problèmes, notamment le verbe qui est souvent polysémique en langue technique.

Cette première méthode est relativement bien connue. Il s'agit de mettre en œuvre certaines techniques développées par la terminologie plus traditionnelle. Cependant, elle n'est valable que pour les termes qui répondent aux critères terminologiques idéaux (sens unique à l'intérieur d'un domaine, appartenance exclusive à ce domaine, etc.). Cependant, même dans le cas d'unités terminologiques « idéales », le codage n'est pas infaillible. Le même texte peut traiter de notions que la terminologie associerait à des domaines différents (dans les faits, il est très difficile d'identifier, pour la plupart des textes techniques, un seul domaine traité).

La seconde méthode de distinction des sens d'une unité lexicale consiste à identifier les liens syntaxiques de cette unité avec d'autres formes linguistiques. Ce second codage permet de contourner certains problèmes entraînés par l'association exclusive de mots isolés dans la base de données lexicales.

Par exemple, le mot *end* se traduit très souvent par *fin* et peut être codé de cette façon dans la base de données lexicales. Cependant, un contexte comme *the end of the cable* se traduira plutôt par *l'extrémité du câble.* On pourra imposer au logiciel les contraintes suivantes :

end suivi de *of* + objet concret = *extrémité de* + objet concret.

Par ailleurs, le verbe *activate*, lorsqu'il est associé à un objet direct comme *function, option, command,* se traduira par *valider* (*une fonction, une option, une commande*) ; cependant, lorsqu'il est associé à *key,* il se traduira par *appuyer sur* (*une touche*).

Soulignons que la syntaxe n'est qu'un moyen détourné d'arriver à certaines fins. Le logiciel de TA reconnaît un lien syntaxique entre deux formes linguistiques ; l'utilisateur a recours à un prétexte pour établir des distinctions sémantiques. Par exemple, dans les deux contextes suivants : *illuminate the indicator* et *the indicator illuminates,* le traducteur reconnaîtra ni plus ni moins le même contexte sémantique (il traduira le verbe de la même façon : *allumer le voyant* et *le voyant s'allume*). Pour le logiciel de TA, il s'agit de deux contextes distincts : dans le premier cas, *illuminate the indicator,* il reconnaît un verbe transitif auquel est associé un objet direct ; dans le second, *the indicator illuminates,* il identifie un verbe intransitif auquel est associé un sujet particulier.

Traitement du verbe en TA

C'est pour l'unité verbale que l'association de formes linguistiques présente les avantages les plus immédiats (bien qu'elle puisse être appliquée à d'autres types d'unités lexicales, notamment les noms dérivés de verbes).

Le verbe *activate,* cité plus haut, peut signifier, dans les textes traitant d'informatique, *activer, lancer, valider* ou *appuyer sur.* La base de données lexicales du logiciel de TA permet de coder un seul équivalent. Le choix pourra porter sur *activer.* Ce premier codage résulterait en une reproduction systématique de *activer* pour toutes les occurrences de *activate,* comme l'illustre l'exemple ci-dessous :

Sortie-machine 1 (entrée *activate* - *activer* dans la base de données lexicales) :

Activate the following function... - Activer la fonction suivante...
The command will be activated once you activate the enter key. - La commande sera activée, une fois que vous activez la touche retour.
To activate the program, activate the F1 key. - Pour activer le programme, activer la touche F1.

Si *activate* est traité différemment et qu'on lui associe des objets directs comme *function, option, command, program* et *key,* en précisant au logiciel de TA que dans les trois premiers cas l'équivalent souhaité est *valider,* dans le cas de *program, lancer* et, dans le dernier cas, *appuyer sur,* le logiciel reproduira ces équivalents lorsque le texte de départ présentera les conditions précisées dans le système. Par exemple :

Sortie machine 2 (*activate* + forme en fonction d'objet direct) :
Activate the following function... - Valider la fonction suivante...
The command will be activated once you activate the enter key. - La commande sera validée, une fois que vous appuyez sur la touche retour.
To activate the program, activate the F1 key. - Pour lancer le programme, appuyer sur la touche F1.

L'identification de liens entre formes linguistiques représente une solution plus appropriée que le découpage des sens en domaines de spécialité pour la distinction des significations multiples d'une forme verbale. Cette dernière remarque s'applique particulièrement aux verbes polysémiques qui ne peuvent être identifiés à un domaine unique. Les exemples suivants illustrent cette dernière remarque.

Le verbe *clear* (employé comme verbe transitif) a des significations diverses que peut expliciter l'objet direct qui lui est associé (*remettre à zéro, effacer, supprimer, vider, nettoyer, éteindre,* etc.). Aucun de ces équivalents ne peut être lié exclusivement à un domaine particulier. Il s'agira, afin d'éliminer ou de réduire les risques d'ambiguïté, de préciser les contraintes suivantes (tableau 3) :

VERBE	OBJET DIRECT	TRADUCTION
CLEAR	a counter, a voltmeter, etc. (a measuring device)	remettre (un appareil de mesure) à zéro
CLEAR	a screen	effacer un écran
CLEAR	an alarm	supprimer une alarme
CLEAR	a memory	vider une mémoire
CLEAR	a filter	nettoyer un filtre
CLEAR	an indicator	éteindre un voyant

Tableau 3.

Le verbe *clear* illustre la transformation requise en langue d'arrivée en fonction d'un objet direct différent. Il s'agit, dans chaque cas, de contextes syntaxiques identiques. Chaque forme précisée au logiciel de TA constitue en fait un contexte sémantique différent.

L'exemple suivant, liste des transformations requises pour le verbe *break* (tableau 4), illustre la diversité des contextes syntaxiques analysables par un logiciel de TA.

Dans les deux premiers contextes (*break a circuit, a connection* et *break the execution of a program,* etc.), *break* est transitif ; la distinction réside dans la complémentation qui lui est associée. Le troisième exemple (*break into oscillation*) illustre un emploi intransitif du verbe lorsqu'une préposition (*into*) et un circonstant (*oscillation*) particuliers lui succèdent en contexte. Les trois derniers exemples illustrent des emplois de *break* assorti d'une particule (*down* et *out*) qui requièrent une transformation particulière en fonction d'un objet, d'un sujet ou d'un circonstant particulier.

Sujet	VERBE	Particule	Objet direct	Préposition	Circonstant	Traduction
	BREAK (V.T.)		circuit connection			couper (circuit, connexion)
	BREAK (V.T.)		execution	of	program routine	interrompre (programme, sous-programme)
	BREAK (V.I.)			into	oscillation	entrer en oscillation
	BREAK (V.T.)	down	list graph			subdiviser (liste, graphique)
machine engine	BREAK (V.I.)	down				(machine, moteur) tombe en panne
	BREAK (V.I.)	out		of	loop	sortir de (boucle)

*VT verbe transitif, *VI verbe intransitif.

Tableau 4.

Dans les exemples précédents, les distinctions sémantiques sont établies en fonction de contextes très restreints (souvent un seul nom). Cependant, ces distinctions peuvent revêtir un caractère beaucoup plus général. Par exemple, le verbe *locate* se traduira fréquemment, en langue technique, par *localiser* lorsqu'il est transitif et utilisé à la voix active (*locate this part in assembly X - localiser cette pièce dans l'ensemble X*). Cependant, lorsqu'il est utilisé à la voix passive en anglais, le verbe se rend par *se trouver* (*this part is located in assembly X - cette pièce se trouve dans l'ensemble X*).

Il s'agira de préciser au logiciel de TA les conditions suivantes :

locate (V.T.) actif + nom (objet direct) *localiser* + nom
locate (V.T.) passif + nom (objet direct) nom + *trouver* (pronominal)

Conclusion

Une meilleure description du comportement des formes lexicales dans le texte peut combler des lacunes que présentent les bases de données lexicales en TA[3]. Les méthodes de codage actuelles impliquent cependant que l'on dresse des inventaires exhaustifs des transformations requises pour certaines unités lexicales en fonction du contexte.

Paradoxalement, ce ne sont pas, dans le cas qui nous préoccupe, les outils informatiques qui font défaut. Certains systèmes de TA offrent la possibilité, bien qu'imparfaite, d'établir des distinctions sémantiques en fonction des liens qu'entretiennent des formes linguistiques en contexte. Les lacunes se retrouvent plutôt du côté des ouvrages plus traditionnels qui passent outre à cette information.

Les ouvrages de référence orientés vers l'usage en langue technique ou scientifique ne traitent que des significations de certaines parties du discours, notamment les noms, ou, s'il tiennent compte de parties du discours variées, ne présentent aucun élément de description syntaxique[4].

Dans l'immédiat, le codage des combinaisons de formes linguistiques, aussi bien pour le choix des combinaisons et la sélection de la traduction appropriée, se fait de manière intuitive ou en fonction de critères de probabilité. Cependant, la reproduction des phraséologismes dans un système de TA, afin d'assurer une transformation de qualité, implique, en premier lieu, le recensement de ces combinaisons en langue technique. Il faudra délimiter des zones d'intérêt (les notions de domaines terminologiques ne sont plus valables pour ce genre de travail), développer des méthodes de recensement qui pourront s'inspirer de modèles déjà élaborés pour la langue générale et, ensuite, reproduire les unités phraséologiques dans le système de traduction.

Références

BÉDARD, C. (1986) : *La Traduction technique : principes et pratique*, Montréal, Linguatech.

Entre nous. Bulletin de traduction technique (1978-1986) : Montréal, Linguatech.

HEID, U. et G. FREIBOTT (1991) : « Collocations dans une base de données terminologique et lexicale », *Meta*, vol. 36, n° 1, pp. 77-91.

L'HOMME, M.-C. (à paraître) : « Management of Terminology in a Machine Translation Environment », *Terminology*.

3. Cependant, puisqu'il s'agit d'un traitement automatique et qu'un logiciel de TA, quel qu'il soit, doit toujours procéder d'une forme linguistique pour effectuer une analyse, l'établissement de liens syntaxiques ne représente qu'une solution partielle à la distinction des significations multiples des mots et des termes.

Par exemple, dans un contexte comme *press the key*, le logiciel reproduira *appuyer sur la touche*, si on lui a précisé ces conditions au préalable. Cependant, dans le contexte *press enter, control, alternate, ctrl*, etc., le traducteur saura qu'il s'agit de noms de touches. Si on n'a pas précisé au système que *enter*, etc. concernait des noms de touches, celui-ci sera incapable d'établir le lien entre *press* et *enter*, etc. et traduira par *presser* suivi du nom).

4. Les lexiques et vocabulaires terminologiques établis selon les règles de l'art ainsi que les banques de terminologie, puisqu'ils sont fondés sur le principe de l'univocité, ne comportent pas de verbes, sinon quelques verbes spécialisés univoques.

Certains dictionnaires techniques généraux proposent différents équivalents pour un verbe anglais, mais ne précisent pas toujours dans quel contexte ces équivalents s'appliquent.

L'HOMME, M.-C. (1993) : « Le verbe en terminologie : du concept au contexte », *L'Actualité terminologique,* 26-2, pp. 17-19.

PHAL, A. (1971) : *Vocabulaire général d'orientation scientifique,* Paris, CRÉDIF.

ROBERTS, R. P. (1993) : « Phraseology: The State of the Art », *L'Actualité terminologique,* 26-2, pp. 4-8.

SADLER, L. et D. ARNOLD (1992) : « Unification and Machine Translation », *Meta,* 37-4, pp. 657-690.

SAGER, J. C. (1990) : *A Practical Course in Terminology Management,* Amsterdam/Philadelphia, John Benjamins.

21

Pour l'analyse des sous-langages en traduction automatique

Graham RUSSELL et Pierrette BOUILLON[1]

ISSCO, Université de Genève, Suisse

• *Abstract* •

Two kinds of strategy can be adopted in order to deal with the current impossibility of high-quality general-purpose MT. The first is to accept low-quality output, with post-editing for subsequent publication, or simply to extract the broad sense of a text. The second, where better performance is essential, relies on measures intended to simplify the problem ; among these are pre-editing or the use of a controlled input language. Applications within restricted domains remain the most feasible.

This article deals with two related issues: that of evaluating the "MT-tractability" of a sublanguage, and that of defining a sublanguage suitable for automatic processing. Different techniques are presented for addressing these questions, taking into account both grammatical properties and the requirements of users and domain experts.

The work described here takes as its starting point a bilingual corpus. It proceeds by identifying sources of complexity, on a number of levels: lexical (ambiguity, synonymy, complex lexical (ambiguity, synonymy, complex lexical units), syntactic (variation, attachment ambiguities), semantic and translational (ambiguity, non-locality of relevant information). Modifications are proposed to make the texts simpler and more systematic, so reducing, and where possible eliminating, this complexity. The objective is to create a sublanguage which meets as closely as possible the requirements for automatic processing, while retaining a maximum of expressivity.

We discuss a number of computational tools, most of which have been devised for other purposes, and examine their application to the relevant tasks: analysis of the source text, domain modelling, establishing the translation relation, and generating the target text. For various reasons, interaction with human experts is envisaged: the methods described make no attempt to "understand" the texts ; factors of style and the relative importance of different parts of the message can only be judged by users ; and potential conflicts arising between alternative simplifications must be resolved using knowledge of the MT system in question.

1. Le travail rapporté ici a été mené dans le cadre du projet Fonds National Suisse No. 12-36505-92. Nous remercions Ted Dunning pour son programme chi2 et Dominique Petitpierre pour ses discussions.

Introduction

Les applications du traitement du langage naturel (TAL) (dont la traduction automatique constitue un cas particulier) exigent du temps et des ressources considérables pour développer les descriptions linguistiques dont dépendra la performance du système. On pourrait, de ce fait, être tenté d'essayer de réutiliser, dans de nouveaux projets ou applications, les grammaires et les lexiques existants. Les recherches dans le domaine des *ressources linguistiques réutilisables* ont d'ailleurs bénéficié d'un grand intérêt ces dernières années (voir, par exemple, Arnold *et al.* 1993). Les résultats ne sont cependant pas encore tout à fait concluants et nous avons de nombreuses raisons de penser que tous les problèmes ne seront pas résolus.

Construire des systèmes de TAL généraux est un objectif séduisant, mais inaccessible. Idéalement, il devrait être possible d'écrire, pour une langue, une grammaire unique et complète qui, couplée à un lexique assez important, nous permettrait de considérer qu'on en a terminé avec la tâche linguistique. En pratique, les choses ne sont pas aussi simples. Sauf exceptions (comme, par exemple, le *Core Language Engine* développé au SRI ; Alshawi 1992), les descriptions linguistiques sont généralement liées à des applications particulières du TAL. La raison principale en est pratique. Généralement, on pare au plus pressé : il est plus important de construire un système performant dans le domaine choisi que d'en élaborer un plus général, susceptible d'être utilisé, avec quelques modifications, dans d'autres domaines. De plus, il faut se rappeler que les applications du TAL ne consistent pas uniquement en descriptions linguistiques : elles impliquent des analyseurs et des générateurs couplés à des bases de données et des mécanismes d'inférences, étroitement liés aux informations linguistiques. Ainsi, lorsqu'un raisonnement lié à un domaine spécifique s'effectue par un programme existant, développé pour traiter des structures de données de type particulier, il peut arriver que les objets produits par l'analyseur et fournis au générateur soient déterminés, non pas par des considérations abstraites concernant les phénomènes linguistiques généraux, mais plutôt par des contraintes d'environnement. Il est aussi possible que les informations contenues dans ces représentations ne soient pertinentes que pour un nombre limité de tâches. Puisque cette interdépendance n'est pas facile à prédire, il faut trouver un moyen de modifier les représentations, soit à l'intérieur de la description elle-même, soit grâce à un interface entre les composantes linguistiques générales et spécifiques au domaine.

Mais une seconde raison, particulièrement intéressante dans le contexte de cet article, rend compte de la spécificité des systèmes : de nombreuses applications sont d'un aspect linguistique relativement simple. Alors qu'une grammaire complète se doit de traiter tous les phénomènes syntaxiques du langage et un lexique des centaines de milliers d'entrées et de sous-entrées, les grammaires développées pour des applications précises ne doivent pas faire face à une telle variété.

Le TAL basé sur les sous-langages offre des avantages évidents : un vocabulaire restreint (qui limite le problème de l'ambiguïté syntaxique et sémantique), un nombre réduit de constructions syntaxiques et une spécification plus complète du domaine du discours. Il permet ainsi un développement plus rapide et de meilleures performances. Cependant, les descriptions des sous-langages, quoique moins décourageantes, peuvent être longues, complexes et difficiles à mener à terme. Différentes approches ont été adoptées, mais il n'en existe pas de parfaite. Ainsi, il serait préférable de ne pas re-

partir à zéro pour chaque description ; mais les raisons mêmes qui rendent les sous-langages si attractifs entravent précisément le transfert des descriptions d'un domaine à l'autre. En effet, le lien étroit qui unit les aspects linguistiques et ceux relatifs au domaine limite les descriptions à un seul domaine du discours et à une seule forme d'expression. Par ailleurs, il n'est pas non plus intéressant de recourir à une description linguistique générale pour le traitement d'un sous-langage spécifique. Il est en effet probable que le sous-langage contient des éléments qui ne sont pas traités dans la description générale et, de plus, il n'est pas utile d'utiliser et de maintenir une description plus complexe que nécessaire.

Supposons que, disposant de grammaires stables et complètes, nous souhaitions en extraire les parties pertinentes pour le traitement d'autres sous-langages. Nous nous heurtons à un premier écueil : pour identifier les parties pertinentes dans ce qui serait inévitablement un très grand ensemble de données, il faut pouvoir comparer les phénomènes textuels aux informations de la grammaire, pour sélectionner les informations nécessaires sur base des propriétés du sous-langage. Ensuite, ces données, extraites de leur description initiale, doivent nécessairement garder leur interprétation. En effet, l'écriture d'une grammaire présume souvent que les informations sont distribuées à l'intérieur d'une description globale de telle façon qu'une information particulière ne fournit qu'un traitement partiel d'un phénomène linguistique, d'autres informations différentes contraignant et complétant les premières, pour former globalement une analyse correcte. Enfin, beaucoup de sous-langages ne sont pas, à proprement parler, des sous-ensembles du langage général ; il est donc tout à fait possible qu'une grammaire générale qui ne tienne pas du tout compte du sous-langage particulier à traiter, ne décrive pas certains phénomènes de ce sous-langage.

Il existe un autre moyen de faciliter la description des sous-langages : définir des techniques d'automatisation partielle du processus. Différentes méthodes sont utilisées pour extraire des textes les informations linguistiquement utiles ; lorsqu'il s'avère impossible de répartir les données dans les différentes applications, à défaut de réutiliser les descriptions elles-mêmes, on peut toujours recourir aux outils qui permettent le développement.

Ces dernières années, un travail considérable a été accompli dans le domaine de l'extraction de différents types d'informations par des méthodes statistiques. Généralement, ces dernières requièrent une grande quantité de textes ; bien que d'importants corpus lisibles sur machine soient disponibles, en plusieurs langues, à l'initiative de, par exemple, l'*ACL Data Collection Initiative,* l'*European Corpus Initiative* ou du *Linguistic Data Consortium*[2], il est peu probable que les sous-langages d'une application donnée soient couverts par l'un d'eux. D'ailleurs nombre d'applications ne disposent que de peu de textes. Le corpus des bulletins d'avalanches qui est à la base du travail décrit ici contient environ 150 000 mots pour l'allemand et autant pour le français. Les corpus de cette taille se prête particulièrement bien au type de travail proposé ici : tout d'abord, les textes d'un sous-langage varient moins que les textes généraux ; les phénomènes caractéristiques des sous-langages apparaissent donc plus souvent que dans le langage général et sont donc susceptibles de fournir un nombre d'exemples suffisants ; ensuite, le processus, qui ne doit pas être entièrement automatique, peut être supervisé par un être humain.

2. Voir Warwick-Armstrong (ce volume) pour un exposé détaillé des initiatives dans ce domaine.

À l'être humain peuvent être dévolus plusieurs rôles. Source de connaissances linguistiques *a priori,* il peut par exemple indiquer, pour un sous-langage en allemand, que les articles varient en genre, nombre et cas. De même, de nombreux aspects de la connaissance relative au domaine peuvent et devraient être traités comme données ; si un sous-langage utilise une terminologie fixe ou un ensemble de schémas phrasaux définis préalablement, ces derniers peuvent être utilisés pour caractériser le sous-langage. Les utilisateurs potentiels d'un système peuvent aussi imposer des restrictions sur la forme d'expression à laquelle le sous-langage doit se conformer : alors qu'un sous-langage simplifié au maximum n'admettra pas plus qu'une manière d'exprimer un message, il se peut qu'il faille multiplier les formes d'expression pour donner aux utilisateurs une impression de variété. Finalement, le type de procédure envisagée ici n'est pas déterministe : dans certains cas, des simplifications incompatibles sont possibles et le développeur doit faire un choix.

Variation des textes

Un problème plus complexe que celui de la taille de la description linguistique concerne le degré de non-déterminisme ; la principale tâche du TAL consiste à choisir entre différents items d'information linguistique (règles, entrées lexicales, etc.), quand le choix n'est pas déterminé. La complexité du TAL provient de la nécessité de gérer différentes alternatives jusqu'au moment où elles peuvent être éliminées. Des applications qui ne requièrent pas de ressources linguistiques générales résolvent ce problème dans une moindre mesure, mais il reste à déterminer la quantité exacte de ressources linguistiques requises par une application. Comme nous l'avons déjà noté, le langage utilisé dans une collection de textes se distingue normalement de ce qui est inhérent à cette application, c'est-à-dire les concepts de l'application et les relations entre eux. Dans la plupart des cas, il devrait donc être possible, en principe, de définir un langage modifié, basé sur les données du corpus de l'application, mais tantôt simplifié, tantôt développé, de manière à l'adapter au traitement automatique.

Dans ce but, nous devons identifier des types de variation qui conduisent à un non-déterminisme et définir des méthodes pour les quantifier. Voici quelques exemples de variations intéressantes pour notre sujet :

Lexicales les synonymes, qui réfèrent de différentes manières à un même concept ou une même relation ; l'ambiguïté catégorielle par laquelle une forme peut jouer plus d'un rôle syntaxique ; l'ambiguïté sémantique, où une forme peut jouer plus d'un rôle sémantique.

Syntaxiques l'ambiguïté structurale, quand, par exemple, la grammaire permet de nombreuses possibilités d'attachement de groupes prépositionnels ; les paraphrases, qui expriment le même sens par différentes phrases – y compris par des variations dans l'ordre des mots.

Traductionnelles les décalages : une expression en T_1 correspond systématiquement à plusieurs expressions en T_2 ; les décalages déterminés par le contexte : une expression en T_1 correspond à plusieurs équivalents en T_2, en fonction des autres informations du texte.

Cooccurrence et tables de contingence

La notion de cooccurrence significative est importante pour les approches statistiques du TAL et sous-tend les recherches décrites dans cet article. Deux questions doivent être résolues avant d'entrer dans le vif du sujet : (1) qu'entend-on par cooccurrence et (2) dans quels cas une cooccurrence est-elle significative, plutôt qu'arbitraire ?

De manière très générale, une occurrence d'un item dans une position particulière à l'intérieur d'un texte peut être considérée comme un type d'événement, tandis que sa non-occurrence dans cette position représente un autre type d'événement. Ainsi, pour un item dans une position, il y a deux possibilités. Si nous nous intéressons au modèle de cooccurrences de deux items distincts, quatre types d'événements doivent être pris en considération : ceux où les deux items arrivent en même temps, ceux où aucun des deux n'arrive, ceux où le premier arrive sans le second et ceux où le second arrive sans le premier. Les combinaisons d'un plus grand nombre d'items peuvent être traitées de la même manière : quand il y a trois items, il y a huit types d'événements distincts. Dans cet article, nous ne nous intéresserons qu'aux cooccurrences de paires d'items.

Supposons que nous nous intéressions au modèle de cooccurrences de deux items, A et B ; leur distribution peut être résumée dans une table de contingence du type de celle de la figure 1.

	A	⌐A
B	*p*	*q*
⌐B	*r*	*s*

FIGURE 1 : Table de contingence 2 fois 2.

Les entrées d'une table de ce type représentent le nombre d'occurrences de quatre types d'événements : p donne le nombre d'événements où A et B arrivent en même temps ; q celui où B arrive sans A et ainsi de suite. Donc, « p + q + r + s » correspond au nombre total d'événements, « p + q » au nombre d'événements-A et « r + s » au nombre de non-événements-B. Ce qui changera d'une application à l'autre, c'est le type des événements considérés. Si nous voulons tirer des conclusions sur l'association de deux mots adjacents m_1 et m_2, les événements à considérer sont des paires de mots (des *bigrammes*) tirées d'un corpus représentatif ; les événements-A seront ceux où le mot m_1 apparaît en premier lieu et les événements-B ceux où le mot m_2 apparaît en second lieu. Le nombre total « p + q + r + s » correspond au nombre total de paires de mots trouvés dans le corpus. Par exemple, la table ci-dessous contient

les données pour la phrase *transport(s) aériens(s),* calculées à partir d'un corpus d'articles de journaux français :

	transport	non-transport
aérien(s)	54	324
⌐ aérien	465	3 440 357

Sur un total de 3 441 200 paires de mots, nous avons trouvé 54 exemples de t*ransport aérien,* 324 exemples où *aérien* n'est pas précédé de *transport,* 465 où *transport* n'est pas suivi de *aérien* et 3 440 357 où d'autres mots que *transport* précèdent d'autres mots que *aérien.*

Exploiter les informations contenues dans les tables de contingence implique l'assignation d'une valeur numérique à la distribution dans son ensemble, de manière à ce que différentes distributions puissent être comparées sur une base uniforme. Souvent, il s'agit d'une valeur de probabilité ; on présume que la probabilité d'occurrence d'un mot peut être obtenue en comparant le nombre de fois où il se produit effectivement avec celui où il pourrait le faire. Dans l'exemple ci-dessus, nous avons trouvé 519 exemples de *transport(s)* sur un total de 3 441 200 possibilités. Sa probabilité d'occurrence est donc d'environ 150 sur 1 million.

Différentes méthodes existent pour combiner et comparer les probabilités. Il est possible de considérer simplement la probabilité d'occurrence de A, étant donné une occurrence de B (la probabilité conditionnelle de A étant donné B). Il est possible aussi de comparer la probabilité qu'ont A et B de se produire ensemble avec celle qu'ils ont de le faire indépendamment l'un de l'autre (l'information mutuelle de A et B). Church et Hanks (1989) utilisent cette dernière mesure pour étudier les collocations en lexicographie, tandis que Gale et Church (1991) appliquent une méthode statistique $\phi 2$, améliorée d'un score-t, pour l'alignement de mots dans des corpus bilingues. Le travail décrit ici exigeait que la mesure adoptée puisse être utilisée à partir de corpus assez petits, où la présence d'un mot individuel est assez basse. Dans cette perspective, Dunning (1993) préconise l'utilisation du *rapport de vraisemblance* (*likelihood ratio* (LR)), une notion abstraite, qui permet de comparer la force de deux hypothèses. Les nôtres sont que deux mots sont liés de manière significative et que leur cooccurrence est aléatoire.

Des valeurs élevées pour ce rapport indiquent une probabilité plus basse que l'association soit accidentelle ; une probabilité plus haute témoigne au contraire d'une association significative. Le bigramme *transport aérien* (vu plus haut) obtient un rapport de 646.20, une valeur plus élevée que la valeur 10.72 obtenue pour le bigramme *transport(s) de(s)/du/d'* :

	transport(s)	⌐transport(s)
de(s)/du/d'	87	408 261
⌐ de(s)/du/d'	432	3 032 420

Pour les bigrammes, ce rapport reflète une relation syntagmatique, souvent de nature collocationnelle ; ici, nous pouvons donc conclure que *transport aérien* a plus de chances d'être une collocation intéressante que *transport de*, qui, dans la plupart des cas, correspond simplement à une complémentation (*transport d'armes*) ou une postmodification (*transports d'urgence*).

D'autres relations peuvent être examinées de la même manière. Ainsi, l'application des cooccurrences à la traduction sera discutée dans la suite de l'article.

Cooccurrences dans la traduction

Un texte et sa traduction constituent une source d'informations implicites sur les équivalences entre deux langues. Nous pouvons en effet présumer que deux textes de ce type expriment le même message, si, pour une expression du texte T1, on trouve une expression équivalente dans le texte T2. Pour localiser l'expression équivalente en T2, il est indispensable d'affiner la correspondance traductionnelle, en identifiant les passages équivalents dans les textes source et cible, un processus connu sous le nom d'alignement. Sur base de ces textes alignés, nous pouvons examiner les paires d'expressions dans les différentes régions et vérifier si la relation qui unit ces expressions est régulière. Une relation d'équivalence traductionnelle est régulière quand la distribution de deux expressions E1 de T1 et E2 de T2 est telle que l'une des expressions n'apparaît jamais sans l'autre dans une région. Cette situation est très rare, mais il est intéressant de calculer l'écart entre la distribution effective et la distribution idéale.

L'exactitude des résultats dépend en grande partie de l'alignement. Si les régions alignées ne correspondent pas à des passages équivalents, les informations dérivées de cet alignement seront de faible valeur. De même, un alignement qui met en relation de petites régions permettra de mieux discerner les correspondances traductionnelles des mots et des phrases.

Différentes méthodes d'alignement de textes parallèles ont été proposées. Certaines, comme par exemple Catizone *et al.* (1989) ; Debili et Sammouda (1992) ; Kay et Röscheisen (1993), utilisent des informations linguistiques sous la forme de données relatives à la distribution des mots. D'autres, comme Brown *et al.* (1991) et Gale et Church (1993), sont basées sur des hypothèses relatives à la taille des régions. Puisque le but de cette recherche est l'étude des distributions de mots dans des régions, l'utilisation d'une des méthodes d'alignement implique que le résultat soit vérifié manuellement ; dans le cas contraire, nous risquerions de reproduire les hypothèses, qui ont conduit à l'alignement original.

Dans notre travail, les bulletins allemands et français émis par l'Institut Fédéral pour l'Étude de la Neige et des Avalanches, entre 1989 et 1992, ont été alignés manuellement. Les textes allemands contiennent approximativement 22 500 mots, les textes français 33 000. L'alignement a permis d'identifier un peu moins de 2 000 régions. Des hypothèses simplificatrices ont été utilisées. Tout d'abord, nous n'avons pas tenu compte de la longueur des régions : elles varient entre 5 et 45 mots dans le texte allemand. Ensuite, nous avons considéré que le nombre total d'événements correspond à la moyenne des longueurs des deux textes (27 862), ce qui suppose que les correspondances une-à-plusieurs sont en proportion égale dans les deux directions, ce qui n'est pas le cas ici[3]. De plus, nous avons traité préalablement le texte pour normaliser les chiffres, ponctuations, flexions, etc.

La suite de cet exposé se fonde sur une vue simple de la correspondance traductionnelle. Considérons M^a et M^f, deux mots provenant des textes allemands et français. Les événements à compter sont les suivants :

1. cooccurrences de M^a et M^f dans la même région ;
2. occurrences de M^f dans une région sans occurrence de M^a ;
3. occurrences de M^a dans une région sans occurrence de M^f ;
4. cooccurrences de mots sans M^a ni M^f.

Si on se réfère à nouveau à la table de contingence de la figure 1, les classes d'événements 1,...4 correspondent respectivement aux nombres p, q, r et s, si A correspond à M^a et B, M^f.

La notion de cooccurrence utilisée ici est légèrement plus complexe qu'on pourrait le croire. Contrairement aux exemples de la section précédente, les possibilités de choix dans la sélection d'une paire de mots dont les membres appartiennent à des régions assez larges sont beaucoup plus élevées. De plus, un mot peut apparaître plus d'une fois dans une région et, dans ce cas, il faut éviter de compter plusieurs fois une simple occurrence de ce mot.

Plus la cooccurrence d'une paire de mots M^a et M^f est fréquente dans une région (plus la valeur de p est élevée) et moins souvent ces mots apparaissent l'un sans l'autre (plus basses sont les valeurs de q et r), plus il est fondé de les considérer comme des équivalents traductionnels. La paire est *parfaite* si les mots apparaissent toujours ensemble et avec une fréquence qui rend peu probable une association par hasard. En réalité, différents facteurs peuvent perturber cette relation parfaite, mais ces facteurs sont en eux-mêmes intéressants dans le contexte de la simplification des sous-langages.

Dans la suite, nous illustrerons le rôle de ces mesures d'association, dans l'établissement de différents types de relations traductionnelles.

Équivalents traductionnels

Dans les cas les plus favorables, nous pouvons coupler des paires de mots allemands et

3. Les noms composés allemands se traduisent typiquement par des syntagmes en français ; voir Bouillon *et al.* (1992) pour des exemples et des discussions.

français sur base de leur position à l'intérieur d'une liste de paires de mots ordonnées en fonction de leur score de LR. La paire <Niederschlag, précipitation>, par exemple, reçoit le score suivant :

allemand	français	LR
Niederschlag	précipitation	484,39
Niederschlag	jour	158,54
niederschlagsfrei	précipitation	137,59

Elle obtient le score le plus élevé, dans le sens où aucune autre paire <Niederschlag, M^f> ou <M^a, précipitation> n'obtient un meilleur score dans la table ci-dessus : les deux paires incluant *Niederschlag* ou *précipitation* ont en effet un score beaucoup plus faible.

De nombreuses paires ne pourraient être extraites si nous appliquions rigoureusement à d'autres exemples ce critère plutôt simple.

allemand	français	LR
Wetter	temps	186,53
Wetter	ensoleillé	92,48
sonnig	temps	264,25
Witterung	temps	110,76
sonnig	ensoleillé	443,82
sonnig	doux	115,16

Dans ce second exemple de paires qui impliquent *Wetter,* la paire <Wetter, temps> obtient le meilleur score. Parmi les autres paires du type <M^a, temps>, l'une, <sonnig, temps>, a un meilleur score ; l'autre, impliquant *Witterung,* peut directement être écartée : son score est sensiblement inférieur. La paire <sonnig, temps> ne doit

cependant pas être considérée comme un équivalent traductionnel, puisque, si nous examinons les paires qui impliquent *sonnig*, nous en trouvons une qui obtient un meilleur score : <sonnig, ensoleillé>, ce qui peut motiver une préférence pour la paire <Wetter, temps>.

Alternatives traductionnelles

Pour que la relation de traduction soit la plus régulière possible, il serait utile de détecter les cas où une expression d'un texte correspond à plusieurs autres dans le texte cible ; comme, parfois, ces variations peuvent être motivées par de bonnes raisons, c'est à un humain de décider si elles sont nécessaires, peut-être le reflet d'une dépendance conceptuelle qui a échappé aux mesures d'association, ou non.

Prenons l'exemple des deux mots allemands (extraits des bulletins d'avalanches) *Gebiet* et *Region* ; tous deux sont couramment traduits par *région*. Les LR sont indiqués dans la table suivante :

allemand	français	LR
Gebiet	*région*	*554,04*
Gebiet	*dans*	*351,94*
Gebiet	*autre*	*304,92*
Gotthardgebiet	*région*	*538,68*
Region	*région*	*535,20*
uebrig	*région*	*531,78*
Gotthardgebiet	*Gotthard*	*1029,84*
Region	*dans*	*370,14*
uebrig	autre	729,11
uebrig	dans	449,95

Les paires <gebiet, région> et <Region, région> ont des scores quasi similaires. Bien que la paire <Gotthardgebiet, région> ait obtenu un score légèrement plus élevé, elle peut vraisemblablement être écartée par le score plus élevé de la paire <Gotthardgebiet, Gotthard>. Comme auparavant, si nous voulons être certains qu'il ne faille pas prendre en considération d'autres paires du type <Region, M^f>, <Gebiet, M^f> et <M^a, région> de score similaire, il faut rechercher les paires qui impliquent M^a et M^f avec des scores plus élevés. Dans ce cas, les paires <uebrig, M^f>, <M^a, autre> et <M^a, dans> ont un meilleur score.

Les erreurs de traduction

Les erreurs de traduction peuvent, dans une certaine mesure, être considérées comme une sous-classe des alternatives traductionnelles, avec la distinction conceptuelle importante que, dans ce cas, il s'agit d'alternatives incorrectes. Pour les besoins de l'exposé, nous allons considérer qu'une erreur potentielle de traduction se reflète dans le score, comme le montre l'exemple suivant :

allemand	français	LR
Visp	Viège	292,56
Goms	Viège	109,16
Visp	Visp	14,63
Verfestigen	Visp	12,32

La paire qui inclut la traduction correcte de *Visp,* <Visp, Viège>, obtient le score le plus élevé pour *Visp* et *Viège.* Il n'y a qu'une occurrence de *Visp* dans le corpus français, le plus fortement associée à l'allemand *Visp.*

Les erreurs de traduction se manifestent parfois de manière moins évidente dans les scores d'association. La paire *Puschlav* et *Poschiavo* entretient une relation plus complexe :

allemand	français	LR
Puschlav	*val*	*54,09*
Puschlav	*Poschiavo*	*38,35*
Puschlav	*Bergell*	*31,62*
Puschlav	*Puschlav*	*18,65*
Puschlav	*reste*	*16,96*
Bergell	*Poschiavo*	*35,344*
Puschlav	Poschiavo | Puschlav	60,82

L'italien *Poschiavo* se traduit en français et en allemand par *Puschlav* ; comme nous pouvons le remarquer, la paire attendue <Puschlav, Puschlav> a cependant un score inférieur à celle de <Puschlav, Poschiavo> et à celui d'autres paires incorrectes (comme <Puschlav, Bergell>).

Nous pouvons néanmoins confirmer que *Poschiavo* et *Puschlav* sont deux traductions alternatives de l'allemand *Puschlav*, en calculant la force d'association obtenue quand les deux formes sont combinées. La notation *Poschiavo|Puschlav* dans la dernière ligne de la table indique une disjonction, dans le sens où l'ensemble des régions prises en considération dans le calcul correspond à l'union des ensembles dans lequel *Poschiavo* et *Puschlav* apparaissent.

Phrases et noms composés

En guise de dernier exemple, nous abordons la question des correspondances une-à-plusieurs. Comme nous l'avons déjà brièvement mentionné plus haut, dans le corpus et, plus généralement, dans d'autres paires de textes allemands et français, les mots simples allemands se traduisent régulièrement par un syntagme en français, comme *Alpennordhang* (versant nord des Alpes) et *Schneebrettgefahr* (danger de plaques de neige). Nous reprenons dans la table suivante les scores LR les plus hauts pour *Alpennordhang* :

allemand	français	LR
Alpennordhang	versant	1946,63
Alpennordhang	nord	1921,46
Alpennordhang	Alpes	1917,35
Alpennordhang	Grisons	1135,01
Alpennordhang	centre	775,39
Alpennordhang	Valais	740,18

L'association entre *Alpennordhang* et différents mots français est très forte et les trois paires aux scores les plus élevés sont précisément les paires valides. Néanmoins, il faut à nouveau déterminer sur quelle base inclure *versant, Alpes* et *nord,* mais pas *Grisons, centre* ou *Valais,* qui, notons-le, ont un score beaucoup plus élevé que la plupart des paires correctes (*Berg* et *montagne* ont, par exemple, un score de 137,03).

Rappelons que les associations $<M^X,M^Y>$ ne prennent pas en compte les similarités dans la distribution des différents mots M^Y1 ... M^Yn dans un texte simple. Si nous dérivons une distribution pour deux ou plusieurs mots, nous pouvons la considérer comme une distribution d'un seul mot. Par exemple, comparons la table où figure la paire <Alpennordhang, versant> avec celle obtenue quand les distributions de *versant* et *nord* sont combinées :

X	Y	X Y	⌐X Y	X ⌐ Y	⌐X ⌐ Y
Alpennordhang	versant	208	58	42	27 554
Alpennordhang	versant & nord	208	13	42	27 599

La distribution pour *versant & nord* est obtenue en prenant l'intersection des ensembles de régions dans laquelle les deux mots apparaissent. Alors que *versant* apparaît 58 fois dans une région, sans *Alpennordhang, versant* et *nord* n'apparaissent que 13 fois sans *Alpennordhang.* Par contraste, les colonnes notées « X Y » et « X ⌐ Y » montrent des chiffres identiques, ce qui reflète l'observation que, dans le contexte de *Alpennordhang*, la présence de *versant* détermine celle de *nord.*

La combinaison des distributions nous permet d'étudier les associations une-à-plusieurs du type de celles impliquées dans les correspondances phrasales. La table suivante donne les LR quand les 6 homologues de *Alpennordhang* aux scores les plus élevés sont intégrés dans le calcul du LR :

allemand	français	LR
Alpennordhang	versant	1946,63
Alpennordhang	versant & nord	2126,60
Alpennordhang	versant & nord & Alpes	2150,56
Alpennordhang	versant & nord & Alpes & Grisons	1266,38
Alpennordhang	versant & nord & Alpes & Grisons & centre	701,17
Alpennordhang	versant & nord & Alpes & Grisons & centre & Valais	306,61

L'intérêt de cette méthode d'analyse est la distinction nette qui apparaît entre *nord* et *Grisons,* juste comme nous le voulions. Tandis que le LR augmente quand les distributions de *Alpes* sont combinées avec celles de *versant* et, à nouveau, quand s'ajoutent celles de *nord,* l'addition de *Grisons* provoque une brusque diminution du score, aggravée par l'inclusion de *centre* et de *Valais.*

Conclusion

Dans cet article, nous avons présenté quelques considérations générales sur les besoins du TAL et insisté sur le rôle des sous-langages et des langages contrôlés. Nous avons soutenu que, dans certains cas, il est possible de *rationaliser* un corpus de textes pour définir un sous-langage attractif du point de vue informatique. Pour illustrer cette thèse, nous avons présenté une méthode d'analyse des correspondances traductionnelles dans des textes bilingues alignés, qui permet d'identifier des relations traductionnelles intéressantes entre les mots. Il est évident que, actuellement, cette méthode doit être considérée comme une heuristique – une aide à l'humain qui désire s'attacher aux complexités superflues dans un corpus de textes parallèles.

Références

ALSHAWI, H. (1992) : *The Core Language Engine,* MIT Press.

ARNOLD, D., T. BADIA, J. VAN GENEBITH, S. MARKANONATOU, S. MOMMA, L. SADLER et P. SCHMIDT (1993) : « Experiments in Reusability of Grammatical Resources », *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics,* pp. 12-20.

BOUILLON, P., BOESEFELDT, K. et G. RUSSELL (1992) : « Compound Nouns in a Unification-based MT System », *Proceedings of the Third Conference on Applied Natural Language Processing,* Trente, pp. 209-215.

BROWN, P. F., J. C. LAI et R. L. MERCER, (1991) : « Aligning Sentences in Parallel Corpora », *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* pp. 169-176.

CATIZONE, R., G. RUSSELL et S. WARWICK (1989) : « Deriving Translation Data from Bilingual Texts », U. Zernik (dir), *Proceedings of the First International Workshop on Lexical Acquisition.*

CHURCH, K. et A. P. HANKS (1989) : « Word Association Norms, Mutual Information, and Lexicography », *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics,* pp. 76-83.

DEBILI, F. et E. SAMMOUDA (1992) : « Aligning Sentences in Bilingual Texts: French-English and French-Arabic », *Proceedings of Coling 1992,* vol. 2, pp. 517-524.

DUNNING, T. (1993) : « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics,* 19(1), pp. 61-74.

FINCH, S. P. (1993) : « Finding Structure in Language », thèse, Université d'Edinburgh.

GALE, W. A. et K. W. CHURCH (1991) : « Identifying Word Correspondences in Parallel Texts », *Proceedings of the DARPA Workshop on Speech and Natural Language,* Pacific Grove, CA, 19th-22nd February, 1991, pp. 152-157.

GALE, W. A. et K. W. CHURCH (1993) : « A Program for Aligning Sentences in Bilingual Corpora », *Computational Linguistics,* 19(2), pp. 75-102.

KAY, M. et M. RÖSCHEISEN (1993) : « Text-Translation Alignment », *Computational Linguistics,* 19(1), pp. 121-142.

22

Traduction interactive : problèmes et solutions (?)

Éric WEHRLI*

Laboratoire d'analyse et de technologie du langage (LATL), Département de linguistique, Université de Genève, Suisse

Introduction

En dehors de cas très particuliers, comme ceux de la traduction de bulletins météorologiques ou de messages d'avalanches, il est généralement admis qu'un système informatique de traduction automatique (TA) ne peut assurer une traduction fidèle sans une intervention humaine. Dans les systèmes classiques de traduction automatique (*SYSTRAN, LOGOS, METAL, GlobalLink,* etc.), cette intervention prend la forme de la postédition, opération lourde et souvent fastidieuse, qui exige les compétences d'un traducteur. En effet, il s'agit de comparer systématiquement le texte source et le texte cible pour déceler les erreurs de traduction et effectuer les corrections nécessaires.

La nécessité de recourir aux services d'un traducteur pour corriger les traductions effectuées automatiquement est due à la complexité de la tâche de traduction. Il n'est guère possible ici de décrire les problèmes multiples que pose la traduction ; mentionnons simplement que le processus de traduction fait intervenir des connaissances extrêmement diverses – connaissances des langues source et cible (lexique, grammaire), connaissances du domaine spécifique au texte à traduire, connaissances générales de nature encyclopédique, capacité de faire des inférences, etc. – que l'on n'est pas en mesure, à l'heure actuelle, de faire apprendre à un ordinateur. Dotés d'une connaissance très partielle des langues concernées, les systèmes de TA ne sont pas en mesure de faire face à toutes les ambiguïtés que peuvent receler les textes qu'on leur soumet. Comme ils fonctionnent de façon autonome, c'est-à-dire de

* Je tiens à remercier tous les collaborateurs du LATL qui ont participé au projet ITS-2, et en particulier Catherine Walther pour ses nombreux commentaires.

manière non interactive, ils sont contraints d'effectuer des choix chaque fois qu'une ambiguïté se présente, mais ne disposant ni des connaissances, ni des mécanismes d'inférence et de compréhension requis pour faire les bons choix, ils sont condamnés à commettre beaucoup d'erreurs. Il faut noter, de plus, que même si le nombre d'erreurs pouvait être ramené à quelques-unes par page, le fait de ne pas savoir où se situent ces erreurs nécessite un recours à une comparaison détaillée des textes source et cible.

On comprend mieux, dès lors, l'orientation prise par la recherche en traduction automatique et en traduction assistée par ordinateur (TA(O)) au cours de ces dernières années. Une des tendances observées est celle de la *traduction basée sur les connaissances,* qui cherche à augmenter les connaissances, et en particulier les connaissances non linguistiques, mises à disposition du système de traduction, de façon à réduire le nombre des erreurs. Dans l'état actuel de l'art en matière de représentation des connaissances, une telle approche n'est guère envisageable en dehors de sous-domaines très limités (voir Nirenburg 1993).

Une autre tendance, beaucoup plus réaliste, est celle de la *traduction assistée par ordinateur (TAO),* qui cherche à améliorer la collaboration entre le système et ses usagers, par le développement de systèmes où la traduction est effectuée par un traducteur, le rôle de l'informatique se limitant à des services, tels que le traitement de texte, la gestion de documents, l'accès à des dictionnaires, à des banques de terminologie ou à des fiches personnelles, etc.

Enfin, une troisième tendance, qui s'est affirmée au cours des quelques dernières années, est celle de la *traduction basée sur des dialogues,* ou simplement *traduction interactive,* qui se présente un peu comme une solution intermédiaire entre TA et TAO. Comme dans les systèmes de traduction automatique autonomes, la traduction est effectuée par le système informatique, et non par un traducteur. Par contre, dans un système interactif, l'intervention humaine ne se situe pas après mais *pendant* le processus de traduction. En effet, l'idée fondamentale sur laquelle repose le concept de traduction interactive est celle de la collaboration active entre système informatique et usager, où ce dernier fonctionne un peu comme un expert, capable d'apporter les informations et les connaissances dont a besoin le programme de traduction. Aux corrections après coup de la postédition on substitue la consultation en cours de processus, lorsque se pose un problème que le système ne peut résoudre seul[1]. Ce type d'approche permet d'obtenir un « premier jet » beaucoup plus élaboré, où le respect du sens est garanti.

L'idée de consulter l'usager en cours de traduction n'est pas nouvelle. Des travaux sur la traduction interactive ont été entrepris il y a plus de vingt ans déjà (voir Kay 1973 ; Melby 1980). Mais, comme le rapporte Blanchon (1994), ces premières expériences n'ont pas dépassé le stade du prototype de laboratoire. Ni les équipements informatiques, ni les technologies linguistiques informatiques ne se prêtaient à ce type d'approche. Cependant, le développement rapide de la micro-informatique et des logiciels « interactifs », ainsi que les échecs répétés des tentatives de traduction automatique autonome (par exemple EUROTRA), ont remis d'actualité cette approche. Il

1. Il va de soi qu'une révision du texte cible reste indispensable, comme, d'ailleurs, dans toute autre forme de traduction. Mais, contrairement à la postédition qui ne peut être accomplie que par des traducteurs, la révision peut être confiée à un rédacteur monolingue.

existe maintenant plusieurs maquettes et prototypes de traduction interactifs, mais, à ce jour, aucun système commercialisé. La réalisation d'un système interactif pose, en effet, des problèmes très complexes. Cet article est consacré à une discussion de certains de ces problèmes et aux solutions que nous avons élaborées (ou que nous envisageons) dans le cadre du développement du système interactif de traduction (anglais-frangais, frangais-anglais) ITS-2 (Wehrli 1993).

Problèmes et solutions (?)

D'un point de vue pratique, l'intérêt majeur que présente l'élaboration d'un système interactif de traduction par opposition aux systèmes non interactifs (ou autonomes) est qu'il ne nécessite pas les compétences d'un traducteur, ni même une connaissance étendue de la langue cible. En effet, on peut raisonnablement penser qu'une clarification suffisante du texte source permet une traduction automatique correcte du point de vue du sens et de la grammaire. Un réviseur en langue cible peut ensuite reprendre le texte cible pour effectuer des corrections stylistiques. Si cette hypothèse se révèle correcte, on notera que les publics visés par les systèmes de traduction autonomes et les systèmes de traduction interactifs sont distincts. Les premiers sont conçus pour des traducteurs (la postédition exige la comparaison systématique des textes source et cible), les seconds pour des rédacteurs monolingues, c'est-à-dire potentiellement tout utilisateur d'un traitement de texte.

Le problème majeur pour la traduction interactive est celui du nombre d'interactions. En effet, peu d'usagers accepteront d'utiliser un système qui engagerait un dialogue de clarification pour tous les mots du texte à traduire. Ce cas de figure peut sembler par trop exagéré, mais le taux d'ambiguïté des langues naturelles est tel que sans précautions particulières, les ambiguïtés d'analyse conjuguées à celles du transfert et de la génération risquent d'atteindre souvent ces limites. La réduction du nombre d'ambiguïtés, et donc du nombre d'interactions, passe par plusieurs mesures, comme l'interaction retardée, l'amélioration de la finesse des analyses syntaxiques et éventuellement aussi l'utilisation de données statistiques.

L'interaction retardée

Un premier pas vers la réduction du nombre d'interactions consiste à éliminer autant que possible les dialogues inutiles, liés à la clarification d'ambiguïtés locales que la suite de l'énoncé parvient à lever. De telles ambiguïtés sont en effet très courantes. Au moment de la lecture (de gauche à droite) du mot *voiles* dans la phrase (1a), l'analyseur ne sait pas s'il s'agit d'une occurrence de *la voile* ou d'une occurrence de *le voile.* Cette ambiguïté disparaît à la lecture du mot suivant, puisque la morphologie de l'adjectif montre qu'il s'agit d'un féminin.

(1) a. Jean regarde les voiles blanches...
b. Les filles du voisin que nous avons rencontré...

De même, dans la phrase (1b), jusqu'à la lecture du participe passé *rencontré,* il ne sait si la relative modifie *filles* ou *voisin.* Dans de tels cas, chercher à clarifier une ambiguïté au moment précis où elle surgit conduit inévitablement à poser des ques-

tions inutiles. Cependant, il serait faux de conclure de cette discussion qu'il suffit d'attendre la fin de l'analyse d'une phrase, voire d'un paragraphe, pour réduire de façon optimale le nombre d'interactions. En effet, à trop attendre avant de commencer à filtrer l'ensemble des analyses, on risque l'explosion combinatoire, c'est-à-dire une prolifération d'analyses telle que le système n'est plus en mesure d'y faire face ! Il faut donc trouver un compromis entre le confort de l'utilisateur et celui du système computationnel. La levée des ambiguïtés ne doit intervenir ni trop tôt, ni trop tard. Le compromis que nous avons retenu pour ITS-2 consiste à retarder les interactions jusqu'au moment où la multiplication des analyses concurrentes dépasse le niveau de tolérance du système informatique. Ce niveau seuil, déterminé de façon empirique, permet généralement d'attendre la fin de la phrase, lorsque celle-ci est relativement courte (moins de 20 mots) ou peu ambiguë.

Du point de vue informatique, la réalisation d'un mécanisme d'interaction retardée nécessite un système relativement complexe de gestion de l'information, de façon à permettre la formulation des dialogues de clarification au moment de l'interaction. Pour ce faire, il est nécessaire de noter, à chaque point d'ambiguïté, toutes les alternatives disponibles et ensuite d'annoter les analyses en fonction des choix dont elles découlent.

Concrètement, dans notre système-prototype ITS-2, ce mécanisme a été implémenté comme suit :

- Chaque point d'ambiguïté est répertorié, avec toute l'information nécessaire pour une formulation de question ultérieure (mots ou syntagmes concernés, type d'attachement, etc.). Chaque alternative est également notée, chacune d'entre elles recevant un numéro d'identification unique.

- Chaque constituant construit par l'analyseur spécifie un ensemble de numéros, correspondant aux choix particuliers dont il est issu.

- Au moment où l'interaction prend place, on parcourt la liste des points d'ambiguïté jusqu'au moment où l'on trouve des constituants issus de choix distincts par rapport à une ambiguïté.

- Un menu avec autant de choix qu'il y a d'options réalisées est alors constitué sur la base des renseignements répertoriés avec les points d'ambiguïté.

- L'usager sélectionne une ou plusieurs lignes du menu. Ces choix correspondent à des numéros spécifiques, qui servent à filtrer la liste des analyses. Seules les analyses qui spécifient les numéros correspondant aux choix de l'usager sont conservées, les autres sont éliminées.

Pour illustrer le fonctionnement de cet algorithme, prenons l'exemple d'une ambiguïté lexicale :

(2) je suis le dernier coureur.

Rappelons, tout d'abord, que l'analyseur lit les mots de gauche à droite et cherche à construire le plus rapidement possible les structures syntaxiques sur la base des mots qu'il a lus. Dans une phrase comme (2), au moment où il parvient au mot

suis, il note la forme ambiguë *suis* et les interprétations possibles, soit, « forme du verbe *suivre* » et « forme du verbe *être* ». Ces deux options reçoivent chacune un numéro d'identification unique, disons 1 et 2. À ce stade, l'analyseur aura construit deux analyses distinctes. Celle construite sur l'interprétation *suis-suivre* comporte le numéro 1 dans l'ensemble de ses choix, alors que celle qui correspond à l'interprétation *suis-être* spécifie le numéro 2. Laissons l'analyseur poursuivre son chemin sans autre interruption, avec les mots *le* (ambigu – pronom clitique ou déterminant défini) et coureur. Arrivé à la fin de la phrase, l'analyseur fournit deux analyses complètes. La procédure d'interaction retardée parcourt la liste des points d'ambiguïté à la recherche d'une ambiguïté pertinente. On dira qu'une ambiguïté est pertinente lorsqu'elle permet de discriminer les analyses, autrement dit lorsque les analyses concurrentes ne sont pas toutes issues du même choix par rapport à ce point d'ambiguïté. Ainsi, *le* dans notre phrase (2) est une ambiguïté non pertinente, puisqu'il est interprété comme déterminant dans les deux analyses. Par contre, *suis* est une ambiguïté pertinente puisqu'elle permet de distinguer les deux analyses. Un menu est alors créé sur la base des informations répertoriées (suis-être/suis-suivre) et affiché à l'écran. L'usager choisit l'une ou l'autre option. Grâce au numéro d'identification des options, on peut facilement filtrer l'ensemble des analyses pour ne garder que celles conformes au choix de l'usager.

L'amélioration des analyses

L'amélioration des analyses linguistiques peut conduire à une diminution du nombre d'interactions[2]. En effet, l'addition de contraintes plus fines que celles que l'on rencontre généralement dans les analyseurs de systèmes de TA permet d'éliminer des ambiguïtés, comme l'illustrent les exemples ci-dessous.

(3) a. Les voiles qu'il souhaite acheter semblent être un peu trop grands/grandes.
b. Les beaux/belles voiles.

Dans l'exemple (3a), le mot *voiles* est ambigu, puisque la morphologie du pluriel ne permet pas de distinguer entre les formes masculine et féminine du mot *voile.* Une analyse syntaxique relativement superficielle – comme celles que l'on utilise actuellement dans les systèmes commercialisés de TA – peut sans doute lever une telle ambiguïté sur la base de l'accord entre le substantif et un adjectif comme dans la phrase (3b), où l'adjectif qualificatif est très proche du substantif. Par contre, lorsque l'accord se fait de manière apparemment non locale, comme dans la phrase (3a), une analyse syntaxique très fine est nécessaire pour garantir et vérifier l'accord entre *voiles* et *grands/grandes.*

Ces accords distants sont plus nombreux qu'on ne le pense. En voici un deuxième exemple, dans le cas de relatives. Ici, c'est l'accord du participe passé qui permet de déterminer sur quel syntagme porte la relative (*femme/président*).

(4) La femme du président, qu'une foule nombreuse a accueilli(e) lors de son arrivée,...

2. Il convient cependant d'être très prudent avec les raffinements de l'analyse linguistique, car ceux-ci peuvent également conduire à un plus grand nombre d'ambiguïtés, l'analyseur opérant des distinctions qu'un analyseur plus grossier ignorerait !

Considérons un exemple d'ambiguïté un peu plus complexe. Admettons, en simplifiant un peu les choses, que le verbe anglais *to remember* donne une traduction adéquate pour le verbe réfléchi *se rappeler*, alors que *to remind* est utilisé pour transcrire le sens non réfléchi de *rappeler*. Bref, admettons les correspondances suivantes :

(5) a. Je me le rappelle ⟶ *I remember it.*
b. Je le lui rappelle ⟶ *I remind him/her of it.*

Sur la base de ces faits, il est particulièrement important que l'analyse linguistique permette de distinguer l'emploi réfléchi de l'emploi non réfléchi d'un verbe comme *rappeler*. Or, étant donné l'ambiguïté des pronoms personnels de 1re et de 2e personne, cette distinction nécessite parfois une analyse détaillée de la phrase.

(6) a. Nous avons promis de le lui rappeler.
b. Nous essaierons de nous le rappeler.
c. Nous vous avons promis d'essayer de nous le rappeler.
d. Nous vous avons persuadé d'essayer de nous le rappeler.

Dans l'exemple (6a), il est facile d'identifier la lecture non réfléchie de *rappeler*, puisque aucun pronom de la classe *me, te, se, nous, vous* n'est associé au verbe infinitif. Par contre, dans (6b-d), le clitique *nous* est présent. En fait, la même proposition infinitive *de nous le rappeler* apparaît dans chacune de ces phrases. Or, il correspond à un emploi réfléchi dans les phrases (b et c) et à un emploi non réfléchi dans la phrase (d). Sur la base des correspondances que nous avons postulées, on traduira *rappeler* par *remind* dans les cas (a et d) et par *remember* dans les cas (b et c). On notera que le recours à l'interaction n'est pas justifié dans ces phrases, puisqu'il n'y a pas d'ambiguïté en ce qui concerne la forme réfléchie ou non réfléchie du verbe. Or, pour éviter des interactions inutiles dans (6b-d), une analyse syntaxique très détaillée est nécessaire, qui permettra de déterminer de façon sûre si *rappeler* est utilisé dans son sens réfléchi ou non.

Des phénomènes semblables s'observent dans les phrases clivées, où un constituant de la phrase enchâssée est extraposé :

(7) a. C'est pour vous rappeler ces décisions que Jean vous a écrit.
b. C'est pour vous rappeler ces décisions que vous aviez fait un procès-verbal de la réunion.

Dans de tels exemples, comme dans les précédents, résoudre l'ambiguïté locale sans recours à l'interaction n'est possible que si l'analyse syntaxique est suffisamment fine pour déterminer le sens précis du verbe *rappeler* (réfléchi/non réfléchi).

La formulation des dialogues

Un autre problème lié à l'approche interactive, de nature très différente de celui de la réduction du nombre d'interactions, est celui de la formulation automatique des dialogues de clarification. En effet, dans chaque cas d'ambiguïté que l'on entend soumettre à l'usager, il s'agit de formuler d'une manière aussi claire et aussi concise que possible des menus qui illustrent les différents sens possibles d'un syntagme donné,

de façon à ce que l'usager puisse effectuer un ou plusieurs choix. La question centrale, ici, est de savoir sous quelle forme ces informations doivent être représentées. Dans le cas d'ambiguïtés catégorielles, le problème n'est pas trop complexe puisqu'on peut simplement afficher la liste des catégories possibles pour le mot ambigu, comme pour le mot *ferme* dans l'exemple (8) ou pour le mot *that* dans l'exemple (9) :

(8) Le pilote *ferme* la porte.
ferme : a- verbe
b- adjectif
c- substantif

(9) *He knows* that *sheep cannot speak.*
that : a- complémenteur
b- déterminant
c- pronom

Par contre, au niveau des ambiguïtés d'attachement, la solution optimale est moins évidente. En effet, comme il est souhaitable que le système soit facilement utilisable par des usagers non linguistes, il n'est guère envisageable d'afficher des structures syntaxiques détaillées. Une alternative à la présentation de structures linguistiques est celle de la paraphrase, c'est-à-dire de la reformulation des différents sens d'un syntagme ambigu. Ainsi, dans le système LMT (*Logic based Machine Translation,* voir Ben-Ari 1988 ; Rimon *et al.* 1991) l'ambiguïté de la portée des adjectifs dans une structure coordonnée comme (10a) donne lieu à la paraphrase (10b), où l'adjectif est distribué dans les deux termes de la coordination.

(10) *Good boys and girls go to heaven.*

a- *Good boys and girls go to heaven*
b- *Good boys and good girls go to heaven*
« les gentils garçons et les (gentilles) filles vont au paradis »

Si la technique de la paraphrase ne semble guère poser de problèmes dans un cas comme (10), sa généralisation pour d'autres cas d'ambiguïtés syntaxiques est moins évidente. Un certain danger existe d'engendrer des paraphrases ambiguës, ou peu claires, risquant d'entraîner de la confusion ou même des erreurs.

Une autre technique, plus simple et moins dangereuse que celle de la paraphrase est celle du parenthésage. Elle est utilisée par exemple dans le système LYDIA-1 (voir Boitet 1993 ; Blanchon 1994) entre autres pour le traitement des ambiguïtés de coordination adjectivale :

(11) Il prend les cahiers et les classeurs noirs.

a- Il prend les cahiers et les (classeurs noirs).
b- Il prend les (cahiers noirs) et les (classeurs noirs).

Toujours dans LYDIA-1, le parenthésage peut être combiné avec la paraphrase, par exemple pour le traitement des attachements de syntagmes prépositionnels :

(12) Marie voit l'homme dans le parc.

a- Marie voit (l'homme dans le parc)
b- dans le parc, Marie voit l'homme

Formulation des dialogues dans ITS-2

Dans le prototype ITS-2, il y a deux niveaux auxquels les ambiguïtés sont particulièrement nombreuses. Il s'agit de l'analyse syntaxique et du transfert lexical, qui seront traités à tour de rôle. C'est au niveau de l'analyse syntaxique que surgissent les ambiguïtés catégorielles et les ambiguïtés d'attachement. Pour les raisons mentionnées plus haut, les dialogues de clarification ne font pas appel à la paraphrase. Il est également apparu que la technique du parenthésage n'est guère appropriée pour certains types d'ambiguïtés, tels que les interprétations de syntagmes extraposés interrogatifs ou relatifs (voir infra), ou encore lorsque les phrases atteignent une certaine longueur. Dans ces conditions, nous avons retenu la solution suivante : afficher la tête du syntagme dont l'attachement est ambigu, et la tête du syntagme modifié par l'attachement envisagé, comme illustré ci-dessous :

(13) Il a frappé le chien avec le collier rouge.
avec le collier rouge : a- complément de *frappé*
b- complément de *chien*

(14) Jean a fait porter le paquet à Paul.
à Paul : a- sujet de *porter*
b- complément de *porter*

(15) À qui avez-vous promis d'écrire ?
à qui : a- complément de *promettre*
b- complément de *écrire*

Ces trois exemples illustrent le traitement d'une ambiguïté d'attachement de syntagme prépositionnel (13), d'assignation de fonction grammaticale (14) et d'interprétation de syntagme antéposé (15).

Au niveau du transfert lexical, se pose le problème de la polysémie et des choix multiples (*temps* ⟶ *time/weather, livre* ⟶ *pound/book,* etc.), particulièrement nombreux lorsque l'on travaille avec un dictionnaire de taille réelle, et non avec un microdictionnaire. Le problème à ce niveau est celui du nombre d'interactions, qui sera traité dans la section suivante, et surtout celui de garantir que les dialogues se fassent en langue source. Pour parvenir à lever de telles ambiguïtés sur la base de la langue source, il est nécessaire de disposer d'informations spécifiques associées aux dictionnaires de transfert, comme des définitions ou des synonymes[3]. C'est sur la base de ces informations, que l'usager pourra choisir en langue source le sens le plus approprié d'un mot ou d'une expression. Les quelques exemples très simplifiés ci-dessous illustrent le type de dialogues que l'on entend proposer :

3. Dans son état actuel, le prototype ITS-2 ne permet pas encore un dialogue en langue source pour ce type d'ambiguïté. Les différentes correspondances sont simplement affichées à l'écran.

(16) Jean a acheté des livres.
livres : a- ouvrage (lecture)
b- monnaie

(17) Ils ont regardé les voiles.
voile : a- (n. fém.) voile de bateau
b- (n. masc.) vêtement

Une telle solution est réalisable, étant donné que l'on peut déterminer toutes les ambiguïtés de transfert dans un lexique bilingue. Cependant, elle nécessite un travail considérable lors de la constitution du lexique bilingue et également lors des inévitables mises à jour du lexique, puisqu'il faudra garantir la présence des informations (définitions, synonymes, etc.) nécessaires pour les dialogues de désambiguïsation en langue source.

La modulation du taux d'interaction

La question fondamentale pour la TA interactive, et celle qui déterminera finalement le succès ou – comme lors des premières tentatives – l'échec de cette approche, est celle du nombre d'interactions nécessaire pour obtenir la traduction d'une phrase. Peu d'usagers, en effet, sont prêts à répondre à une avalanche de questions pour traduire une phrase d'une dizaine de mots !

Même si le recours à un analyseur puissant et à l'interaction retardée permet de réduire le nombre d'interactions et d'éliminer les interactions inutiles, le nombre d'ambiguïtés pour un système dépourvu de connaissances autres que grammaticales est susceptible de rester très élevé, notamment au niveau du transfert lexical, et sans aucun doute trop élevé pour certaines applications ou certains usagers. Il est donc souhaitable qu'un système interactif de TA soit pourvu d'un mécanisme de choix *par défaut,* et offre la possibilité de moduler le taux d'interaction, de façon à ce que l'usager puisse choisir le mode de travail le plus approprié pour une tâche de traduction donnée. Pour le système ITS-2, nous envisageons toute une palette de choix et d'options, allant du mode totalement interactif – toutes les ambiguïtés sont soumises à l'appréciation de l'usager – au mode totalement automatique – le système opte pour les valeurs par défaut pour toutes les ambiguïtés. Entre ces deux extrêmes, il convient de trouver le meilleur compromis entre intervention humaine et choix par défaut, en fonction de la qualité souhaitée de la traduction et du temps que l'usager est prêt à consacrer à cette tâche.

Dans cette perspective, les choix par défaut doivent être effectués avec beaucoup de soin. Dans ITS-2, ils sont définis sur la base d'heuristiques et de données statistiques. Par rapport aux ambiguïtés d'attachement, un certain nombre d'heuristiques relativement simples ont été développées sur la base de données psycholinguistiques. Ainsi, on donnera la préférence à un attachement de complément par rapport à un attachement d'ajout. En dehors de tout contexte, la valeur par défaut du syntagme prépositionnel *de la fenêtre* dans l'exemple (18) sera donc complément du verbe *parler* et non circonstanciel de lieu.

(18) Jean parle de la fenêtre.

Pour d'autres types d'ambiguïté comme les ambiguïtés catégorielles et les ambiguïtés de transfert lexical, nous comptons utiliser des données statistiques (fré-

quence) pour déterminer les valeurs par défaut. Il est clair, en effet, que certaines lectures de mots sont beaucoup plus fréquentes que d'autres. Par exemple, le mot orthographique *président*, qui peut être un substantif (*le président français*) ou une forme verbale (*ils président*), est beaucoup plus fréquent dans son premier emploi que dans le second[4]. L'anglais, langue dans laquelle les ambiguïtés lexicales sont particulièrement nombreuses – dues sans doute à la pauvreté de la morphologie – fournit d'excellents exemples d'homographes dont les fréquences sont fort différentes. Ainsi *man, water, flower, tank, fish, wind*, etc. offrent tous une lecture nominale et une lecture verbale, mais la seconde est beaucoup moins courante que la première. Dans des contextes syntaxiques où l'ambiguïté subsiste on utilisera les fréquences soit pour déterminer une valeur par défaut, en mode non interactif, soit pour ordonner les choix dans un menu, en mode interactif. Le même mécanisme s'applique au niveau du transfert lexical, puisque là aussi, en dehors de contextes particuliers, la correspondance *livre* ⟶ *book* est beaucoup plus fréquente que *livre* ⟶ *pound*, et la correspondance *ciel* ⟶ *sky* plus fréquente que *ciel* ⟶ *heaven*. Ces derniers exemples montrent d'ailleurs qu'à côté des fréquences et des heuristiques psycholinguistiques, il serait souhaitable de pouvoir prendre en considération la spécificité du domaine traité ainsi que les choix antérieurs effectués par l'usager. Mais ceci reste musique d'avenir...

Références

BEN-ARI, D., BERRY, D. et M. RIMON (1988) : « Translational Ambiguity Rephrased », *Proceedings of the 2nd International Conference on Theoretical and Methodological Issues in Machine Translation*, Pittsburgh, June 1988.

BLANCHON, H. (1994) : *Lidia-1 : Une première maquette vers la TA.*

BOITET, C. (1993) : « La TAO comme technologie scientifique : le cas de la traduction automatique fondée sur le dialogue », P. Bouillon et A. Clas (dir), *La Traductique*, Montréal, Les Presses de l'Université de Montréal, AUPELF/UREF, pp. 109-148.

BOUILLON, P. et A. CLAS (dir) (1993) : *La Traductique*, Montréal, Les Presses de l'Université de Montréal, AUPELF/UREF.

KAY, M. (1973) : « The Mind System », *Courant Computer Science Symposium 8: Natural Language Processing*, Algorithmics Press, pp. 155-188.

MELBY, A., SMITH, M. et J. PETERSON (1980) : « ITS: An Interactive Translation System », *Proceedings of Coling 80*, pp. 424-429.

NIRENBURG, S. (1993) : « L'interlangue et le traitement du sens dans les systèmes de traduction automatique », P. Bouillon et A. Clas (dir), *La Traductique*, Montréal, Les Presses de l'Université de Montréal, AUPELF/UREF, pp. 91-108.

RIMON, M., MCCORD, M., SCHWALL, U. et P. MARTINEZ (1991) : « Advances in Machine Translation Research in IBM », *Machine Translation Summit III*, pp. 11-18.

WEHRLI, E. (1993) : « Vers un système de traduction interactif », *La Traductique*, Montréal, Les Presses de l'Université de Montréal, AUPELF/UREF, pp. 423-432.

4. Cet exemple est très théorique, car il est difficile d'imaginer une phrase dans laquelle le contexte grammatical ne permettrait pas de lever cette ambiguïté.

23

Topicalisation et focalisation dans un système génératif

Jacques LEROT

Projet GENESE[1], Institut de Linguistique UCL, Louvain-la-Neuve, Belgique

• *Abstract* •

The GENESE Project explores the multilingual generation of sentences from semantic representations as conceptual networks.

A peculiar problem generation has to cope with is that of the communicative perspective induced by topicalization and focalization. This communicative perspective can indeed be expressed by means of a modification of the order of the constituents or by various other means.

In this paper, we first propose a clarification of the very notions of topicalization and focalization.

We then show how the propositional contents expressed in the communicatively neutral semantic representations can be topicalized and/or focalized so as to introduce the adequate communicative perspective. Syntactic projection rules finally operate on such topicalized structures to produce the functional structures and their various realizations.

1. Projet financé par la *Communauté française de Belgique* sous le n° 89/94-137.

La recherche entreprise dans le cadre du projet GENESE vise à développer un système de génération de phrases françaises et néerlandaises à partir de représentations sémantiques. Plus particulièrement, le projet consiste :
- à constituer un corpus de représentations sémantiques de portions cohérentes de textes généraux ;
- à garantir la cohérence des représentations sémantiques par un système de règles de bonne formation organisées dans deux ressources supralinguistiques à vocation universelle :
  - un lexique conceptuel, qui a pour objet d'analyser et de définir les types conceptuels (classes) et de fournir les données utiles pour la paraphrase, et
  - une grammaire conceptuelle, qui fait l'inventaire des fonctions conceptuelles et les définit en termes de fonctions primitives ;
- à développer les règles qui transforment les représentations sémantiques en structures syntaxiques.

En outre, le projet vise à construire un générateur à base sémantique qui réalise la synthèse des éléments pertinents pour l'établissement des structures syntaxiques, c'est-à-dire :
- du contenu sémantique représenté sous forme de réseaux sémantiques ;
- de la perspective communicative, comprenant les opérations de topicalisation et de focalisation ;
- de l'univers du discours et des bases de connaissance ;
- des informations lexicales et grammaticales spécifiques (idiosyncrasiques), etc

Traduction et génération multilingue

Le développement des systèmes de traduction automatique ou semi-automatique se heurte aux limites des modèles linguistiques utilisés et dont quelques-uns sont relativement anciens. Les systèmes de transfert actuellement proposés offrent des solutions ponctuelles difficilement réutilisables pour d'autres applications.

Un système de traduction (semi-) automatique par pivot sémantique, tout irréaliste qu'il apparaisse à l'heure actuelle, offre, d'un point de vue théorique, l'avantage :

- de simuler les opérations intellectuelles du traducteur ;
- de ne pas construire la structure de la phrase cible sur le moule de la phrase source ;
- de convenir pour la traduction entre paires de langues appartenant à des familles totalement différentes ;
- de permettre à la fois la traduction et la génération multilingue et, enfin
- de garantir une réutilisation optimale des ressources utilisées.

De l'avis des observateurs qualifiés, un tel projet n'est pas réalisable sans un développement préalable de la recherche fondamentale en linguistique.

Le problème

Un problème déroutant

Parmi les problèmes fondamentaux à résoudre, la topicalisation et la focalisation, tous deux regroupés sous le terme générique de « perspective communicative », comptent parmi les plus ingrats et plongent les linguistes de tous bords dans la perplexité.

La perspective communicative décrit des fonctions qui ne sont perçues de façon optimale que dans les échanges où les interlocuteurs se trouvent face à face et où il importe au locuteur de se faire comprendre de son interlocuteur et d'amener celui-ci à réagir dans le sens souhaité. Or, ces fonctions sont partiellement neutralisées dans la communication écrite typique où le scripteur, ne connaissant pas son lecteur, se trouve dans l'incapacité de gérer les connaissances de celui-ci et donc de développer une stratégie interactive. Malheureusement, bien qu'une langue soit un code essentiellement oral, la tradition linguistique, suivie en cela par la linguistique computationnelle, accorde une importance préférentielle à l'étude de la langue écrite. Ceci explique à la fois le désintéressement chronique envers la perspective communicative et l'énorme difficulté à aborder ces problèmes à partir de documents écrits.

Diversité des formes d'expression

Les difficultés qui entourent la perspective communicative n'ont rien de surprenant lorsqu'on examine ses diverses formes d'expression.

(a) L'intonation et l'accentuation sont, dans la langue parlée, les instruments les plus efficaces de mise en perspective de l'information.

(b) L'ordre des mots est, dans de nombreuses langues et particulièrement dans les langues qui disposent d'un système de cas élaboré comme les langues slaves, un procédé à la fois simple et fonctionnel.

(c) Les particules, les tournures morpho-syntaxiques et les constructions syntaxiques contribuent elles aussi à mettre le discours en perspective. Le japonais possède un système élaboré de particules à valeur communicative (la particule topicale *wa* et la particule *ga* signalant une information nouvelle). Le français parlé fait un usage fréquent de présentatifs : *il y a... que/qui* et *c'est... qui/que* et pratique volontiers l'extraposition.

Une inflation terminologique

Nous ne pouvons que constater l'inflation terminologique qui entoure la perspective communicative. Certains opposent un *topique* (*topic*) à un *propos* ou à un *commentaire* (*comment*) ; d'autres opposent un *thème* à un *rhème* ou un *topique* à un *focus* (Hockett 1958 : 191 ; Chafe 1975 ; Reis 1977, etc.). Un important courant de recherche utilise la dichotomie « thétique-catégorique » (Kuroda 1972 ; Sasse 1987). Dans l'École de Prague, les oppositions binaires sont remplacées par un système de valeurs scalaires baptisé « dynamisme communicatif » (Firbas 1964 ; Combettes 1983).

Cette inflation terminologique s'accompagne d'une hétérogénéité dans les définitions. Sans entrer dans les détails, nous pouvons distinguer trois grands types de définitions :

(a) morpho-syntaxiques : le topique désigne la première position dans la phrase et le focus le syntagme portant l'accent principal ;

(b) logiques : on oppose « ce à propos de quoi on dit quelque chose » à « ce qu'on en dit » ou une proposition « assertée » à une proposition « présupposée » ;

(c) communicatives : on oppose l'ancien au nouveau, le connu à l'inconnu.

Pour plus de détails concernant ces termes, leurs définitions et leur histoire, on consultera Beneš (1973) et Weigand (1979).

Essai de clarification

Les niveaux de description

Ce premier examen fait apparaître que les confusions terminologiques sont liées au mélange de critères syntaxiques, sémantiques et pragmatiques. Il suscite une première question : quelle est la véritable nature de la perspective communicative ?

Selon Chafe (1975), la perspective communicative n'est pas un phénomène de contenu sémantique, mais de transmission de ces contenus, un phénomène d'emballage (*packaging phenomenon*). Selon Weigand (1979 : 180), l'opposition entre ce dont on parle et ce qu'on en dit n'est ni sémantique, ni syntaxique, mais textuelle. Ceci nous invite à distinguer trois niveaux de description :

(a) un **niveau propositionnel**, où sont rassemblées les données sémantiques brutes, c'est-à-dire un ensemble de virtualités considérées indépendamment de la perspective communicative et de l'univers du discours ;

(b) un **niveau assertionnel**, où les données sémantiques virtuelles issues du niveau propositionnel sont transformées en texte, c'est-à-dire organisées en vue de la communication en fonction des intentions de l'émetteur, des connaissances qu'il présume chez le destinataire et du contexte d'énonciation, etc. ;

(c) un **niveau syntaxique**, où les structures assertionnelles mises en perspective communicative sont exprimées par des unités lexicales, grammaticales et syntaxiques.

Divers auteurs utilisent trois niveaux de description comparables aux niveaux précités. Leurs contenus et leurs appellations présentent toutefois d'importantes différences (Zaefferer 1991). Souvent, on identifie un niveau proprement sémantique. Dans le cas présent, le niveau propositionnel ne contient que les significations dites « locutoires » tandis que le niveau assertionnel comprend en outre les significations dites illocutoires (Jacobs 1984 : 25).

La perspective communicative apparaît comme un phénomène d'organisation de la masse sémantique en vue de la communication et, de ce fait, relève typiquement du niveau assertionnel.

Le connu et l'inconnu

L'opposition entre l'ancien et le nouveau ou entre le connu et l'inconnu est incontestablement la plus populaire des dichotomies traditionnelles, mais aussi la plus problématique. Dans une phrase comme : *C'est Jean qui a fermé la porte*, on peut affirmer que *Jean* constitue un élément nouveau et donc présumé inconnu. Cependant, *Jean* est une personne présumée connue des interlocuteurs. Ainsi posé, le problème est insoluble.

Nous croyons utile de distinguer ici deux notions souvent confondues. D'une part, les **objets individuels** sont connus des interlocuteurs lorsqu'ils font partie de l'univers du discours, c'est-à-dire du contexte d'énonciation et des connaissances dont disposent les interlocuteurs. D'autre part, les informations sont des **propositions** qui, par nature, sont porteuses d'une valeur de vérité. Elles peuvent faire l'objet d'une affirmation, d'une dénégation, d'une question ou être présupposées.

Dans la phrase citée : *C'est Jean qui a fermé la porte*, Jean est une personne individuelle connue des interlocuteurs tandis que l'information représentée par *c'est Jean* fait l'objet de l'assertion. Cette information est donc à considérer comme nouvelle et donc présumée inconnue de l'allocutaire. Il ne faut donc pas confondre : (a)

les éléments connus parce qu'ils font partie de l'univers du discours et (b) la gestion des informations dans le discours. On lira à ce propos les clarifications apportées par Haftka (1978) et Weigand (1979).

Un double système d'opposition

L'imbroglio qui entoure la perspective communicative s'explique également par la confusion de deux systèmes d'opposition. Selon Halliday (1967), il faut distinguer, d'une part, la « structure thématique » opposant ce dont il est question à ce qu'on en dit, et, d'autre part, une « structure informative » opposant une information ancienne à une information nouvelle. Une distinction analogue est faite par Jacobs (1984) qui oppose l'articulation focus/arrière-plan à l'articulation topique/commentaire. La différence entre ces deux types d'opposition peut être illustrée par certaines phrases où les deux articulations coexistent.

(Comment s'appelle ta maman?)
Ma maman, Monique, elle s'appelle.
└ topique ┘ └——— propos ———┘
└ focus ┘ └ arrière-plan ┘

Jean, c'est lui qui a fermé la porte.
└top┘ └——— propos ———┘
└focus ┘ └—— arrière-plan——┘

La topicalisation

Définition

Les informations ne sont assimilées que si elles sont mises en rapport entre elles et avec les connaissances dont on dispose déjà. En effet, une connaissance isolée est aussitôt oubliée. Une bonne gestion de l'information suppose qu'on puisse la stocker en mémoire. Ceci nécessite l'usage de mots-clés ou **topiques** permettant à la fois d'accéder aux informations préexistantes et de ranger les connaissances nouvelles. La reconnaissance du contenu suppose que les topiques ou mots-clés apparaissent clairement (Reinhard 1981). Le topique a pour fonction d'annoncer ce dont il est question dans le message. On peut le considérer comme la rubrique sous laquelle on peut ranger le propos. La topicalisation est le procédé par lequel un émetteur fait connaître le topique de son propos.

Le topique revêt une importance primordiale pour l'interprétation.

(a) Il signale le domaine de connaissance dont le texte relève et active le présavoir des récepteurs. C'est à la lumière de ce présavoir que le message sera interprété.

Un topique comme *le procès* suggère un ou plusieurs accusés, un ou plusieurs

accusateurs, un chef d'accusation, un juge, une sentence, etc. Bref, le topique met en éveil les connaissances relatives au domaine juridique.

(b) Le topique conditionne toute l'architecture du message. Dans un dialogue, les interventions des participants doivent être conformes à l'orientation topicale du texte (principe de coopération de Grice (1975)).

(c) Le topique fournit également la rubrique permettant de ranger les informations nouvelles en mémoire.

Le topique est souvent implicite. Lorsqu'un événement quelconque attire l'attention des interlocuteurs, on s'attend à ce qu'on en parle. L'événement extérieur constitue un topique et ce qu'on en dit est le propos.

On entend un bruit de sifflement provenant de la cuisine. Le locuteur, connaissant l'origine du bruit, présume que son interlocuteur s'interroge à propos de ce bruit. Il dira tout naturellement : *C'est l'eau qui bout.* Cette phrase n'a pas de topique explicite. On peut toutefois expliciter le topique sans difficulté : *Ce bruit, c'est l'eau qui bout.*

Comme on peut ranger une même information sous différentes rubriques, par exemple *le tremblement de terre de Lisbonne en 1755* sous les rubriques *tremblement de terre, Lisbonne, 1755,* on peut topicaliser plusieurs fois un même propos. La topicalisation est donc un phénomène récursif. La topicalisation multiple s'observe aisément dans la langue parlée : *Jean, ses examens, il les a enfin réussis.*

Lorsqu'une phrase est connectée à une phrase dominante, elle n'a généralement pas de topique, car celui-ci se trouve dans la phrase dominante.

Identification

(a) Le topique est reconnaissable au moyen d'un test qui consiste à faire précéder le segment par *à propos de* ou *quant à* ou encore à faire précéder la phrase d'une question du type *Qu'est-il arrivé à x ?, What about x ?* (Gundel 1977 : 32 ; Lakoff 1971 : 236). Les problèmes relatifs aux tests sont discutés dans Weigand (1979 : 177).

(b) Un autre critère de reconnaissance est la faculté d'isoler le topique complètement de la phrase (Altmann 1981 ; Jacobs 1984 : 47).

Des trucs pareils pour les mômes, franchement, qu'est-ce qu'on ne voit pas !
La cantine, on n'a pas à se plaindre (Gadet 1989).

(c) Le topique reçoit un accent pratiquement équivalent à celui du domaine focal ou *focus.*

(d) Le topique a souvent une intonation ascendante dont la fonction est d'annoncer une suite.

(e) Généralement, le topique précède le propos.

Types de topiques

On distingue plusieurs types de topiques.

(a) Lorsque le propos ne peut être accroché à un élément présent à l'esprit, le locuteur crée un topique (**topique d'ouverture**). Dans la langue parlée, la tournure *il y a... que/qui* est utilisée à cette fin.

Il y a un type bizarre qui t'attend en bas.

(b) Le **topique de liaison** sert à accrocher un nouveau propos à une information présente à l'esprit. C'est notamment le cas lorsqu'une phrase apporte une information concernant un élément mentionné dans la phrase précédente.

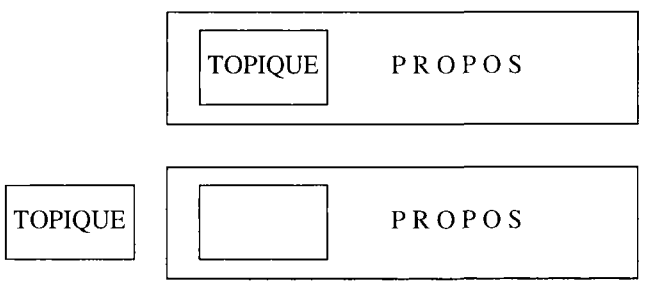
*La mère Michel a retrouvé son chat. **Il** était enfermé dans la remise du boucher.*

(c) Le **topique de contraste** sert à introduire un élément qui contraste avec un élément précédent à propos duquel l'émetteur fournit une information opposée.

(Jean est souvent malade) *Pierre, lui, est resplendissant de santé.*

Le propos ou commentaire

Le propos ou commentaire représente simplement l'information qui a donné lieu à la topicalisation. Le topique est un élément d'information faisant partie du propos. La topicalisation consiste à extraire le topique du propos en y laissant éventuellement une trace pronominale ou adverbiale. C'est pourquoi, le propos se présente souvent comme ce qui reste de l'information après l'extraction du topique.



La focalisation

Illocution et modalisation

L'interprétabilité d'un texte suppose que le récepteur reconnaisse l'objectif poursuivi par le locuteur afin de s'y conformer ou, le cas échéant, de s'y opposer. On appelle **illocution** ou **acte illocutif** l'acte par lequel le locuteur manifeste son intention communicative envers son interlocuteur. L'illocution est une propriété de tout acte com-

municatif. Elle comprend les types bien connus : représentatif ou assertif, interrogatif, promissif, expressif, déclaratif.

L'illocution est généralement modalisée de diverses façons ; atténuée ou renforcée, accompagnée de marques d'égard ou de contraintes, etc.

*Toto, mange **bien** ta soupe.*
*Tu vas manger ta soupe, **ou quoi** ?*
***Et si** tu mangeais ta soupe ?*

Le domaine focal et le domaine scénique

En posant un acte communicatif, le locuteur détermine le contenu qu'il désire introduire dans la conscience de son interlocuteur. Ce contenu, appelé **domaine focal** ou **focus,** ne correspond pas nécessairement à une information nouvelle.

Le locuteur peut focaliser la totalité du propos ou une partie seulement. Dans ce cas, l'information focalisée (le domaine focal) est mise en évidence par rapport aux informations qui se trouvent à l'arrière-plan de la communication.

Comparons :
(1) *Jean est venu hier.*
(2) *Jean est venu HIER.* (accentuation forte sur *hier*)
(3) *C'est hier que Jean est venu.*

En (1) la totalité de l'information est portée à la connaissance de l'interlocuteur, mais dans (2) et (3), l'information est focalisée sur le moment de cette venue, c'est-à-dire *hier* tandis que *Jean est venu* représente l'arrière-plan communicatif.

Les contenus qui constituent l'arrière-plan du message représentent un ensemble de connaissances préalables ou contiennent des explications jugées nécessaires à la bonne compréhension des contenus focalisés. Ces informations constituent une espèce de décor cognitif ou **domaine scénique** dans lequel se situent les informations focalisées. Souvent, le domaine scénique représente des présuppositions pragmatiques, mais ce n'est pas toujours le cas. On évitera d'assimiler la paire assertion/présupposition à l'opposition domaine focal/domaine scénique (Reis 1977 : 217-228).

Plusieurs tests faciles à réaliser permettent l'identification du domaine focal.

(a) Le domaine focal constitue le reste minimum obligatoire sans lequel la phrase est, d'un point de vue communicatif, absurde.

À la question : *Quand Jean reviendra-t-il ?,* on peut donner une réponse élaborée : *Jean reviendra la semaine prochaine pour rechercher quelques documents* ou une réponse succincte : *La semaine prochaine,* mais on ne pourra pas répondre : **Il reviendra pour rechercher quelques documents* ou **Il reviendra.*

(b) Le domaine focal contient l'accent principal de la phrase. Certaines phrases ont un

accent, d'autres en ont deux. Un de ces accents tombe sur le domaine focal, le second est lié à la présence d'un topique (Sasse 1987 : 521).

The SUN is shining.
(Le topique est implicite : on s'interroge à propos du temps.)

HARry's stopped SMOKing.
(Un accent tombe sur le topique *Harry* et le second sur le focus.)

(c) La négation totale, les modalités de type épistémique, les particules modales ne portent que sur le domaine focal (Jacobs 1984).

(d) C'est à l'intention communicative et à l'information contenue dans le domaine focal que réagit l'interlocuteur. Si la réaction doit porter sur une information extérieure au domaine focal, on la signale généralement par *mais.*

Dans : *Jean viendra sans doute avec sa femme. – Je ne le crois pas.* La réaction portre évidemment sur l'affirmation *Jean viendra avec sa femme* et non sur *Jean a une femme.* Si on conteste cette dernière information, on dira par exemple : *Mais je croyais qu'il n'était pas marié.*

L'identification du domaine scénique se fait négativement par référence au domaine focal. Les segments de phrase qui représentent le domaine scénique sont dépourvus d'accent ; ils se placent volontiers avant le domaine focal ; on peut les supprimer sans nuire à la cohérence du discours, etc.

L'acte de prédication

Le domaine focal est un contenu propositionnel dont on peut affirmer la vérité, souhaiter la réalisation, s'informer de la réalité, etc. Il a donc nécessairement un statut relationnel (Jacobs 1984 : 29 ; Weigand 1979 : 173). En produisant un acte de langage, le locuteur pose non seulement un acte locutif et un acte illocutif (Austin 1962), mais également un acte prédicatif, qui, selon J. Searle, fait partie de l'acte propositionnel (Searle 1970 : 24).

L'acte de prédication (au sens pragmatique du terme) consiste à présenter un fait comme une propriété attribuée à une entité. L'entité à laquelle une propriété est attribuée sera appelée « argument » ou « base de prédication » (Sasse 1987 : 555) tandis que la propriété qui lui est attribuée sera appelée « prédicat » (au sens pragmatique du terme). L'entité qui sert d'argument (de base de prédication) est un élément autonome dont l'existence a été préalablement établie.

On peut distinguer plusieurs types de prédication :

(a) la prédication centrée sur un objet connu, c'est-à-dire dont l'argument est un objet, par exemple : *Pierre est malade* ;

(b) la prédication centrée sur un événement, c'est-à-dire dont l'argument est un événement et le prédicat exprime la réalité spatio-temporelle de cet événement, par exemple :

Il pleut signifiant : « pluie (=argument) en cours maintenant (=prédicat) » ;

(c) la prédication centrée sur un objet nouveau, non présent dans l'univers du discours ; dans ce cas, le locuteur doit préalablement poser un acte établissant l'existence de cet objet, par exemple : *Il y a un monsieur qui veut vous parler.*

Lorsqu'un locuteur ne peut assumer totalement la vérité de ses dires, il doit le signaler sous peine de se voir contredit par son interlocuteur. Les modalités épistémiques expriment le degré de croyance du locuteur relativement à ce qu'il affirme. L'acte prédicatif est ainsi modulé selon ce type de modalité dont les valeurs sont la nécessité, la possibilité, la probabilité, la supposition, le doute, etc.

Le système génératif

Difficultés

À la différence des systèmes d'analyse qui procèdent à partir d'une source préexistante (un texte source), un système génératif doit, avant de pouvoir fonctionner, élaborer son propre *input*. Le pouvoir génératif du système est directement lié à la qualité de l'entrée sémantique.

Dans un système génératif, le problème de la polysémie est inexistant, car l'entrée est par nature libre de toute ambiguïté. Par contre, un problème de synonymie surgit, car une même entrée sémantique fournit généralement des formes d'expression multiples. Or, les variations de perspective communicative sont de nature à engendrer de multiples phrases à partir d'un même contenu propositionnel.

César a conquis la Gaule.
César, il a conquis la Gaule.
César, la Gaule, il l'a conquise.
La Gaule, César l'a conquise.
C'est César qui a conquis la Gaule.
La Gaule a été conquise par César.
C'est par César que la Gaule a été conquise.
César, lui, a conquis la Gaule.
César, c'est lui qui a conquis la Gaule.
etc.

Caractéristiques générales d'un système génératif

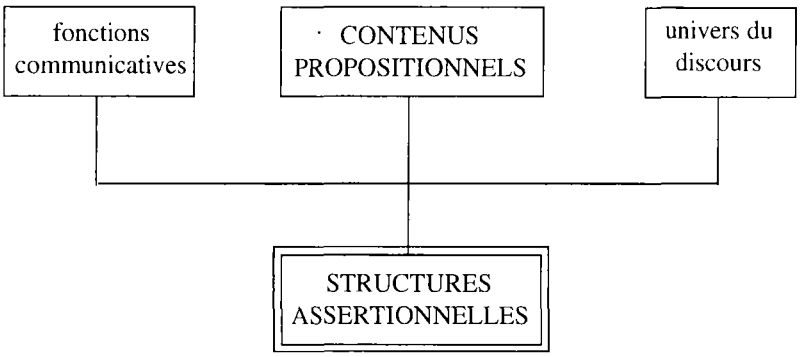
Modèle stratificationnel

Les opérations de projection de données assertionnelles sur des séquences linéaires, c'est-à-dire sur des phrases et des séquences de phrases, sont complexes. En outre, le modèle doit être en mesure de permettre la génération dans des langues de structures très différentes. Nous pensons qu'un modèle de type stratificationnel est susceptible de rencontrer cet objectif, car dans un tel modèle, les opérations de projection s'ef-

fectuent par paliers successifs. Chacun de ces paliers correspond à un **plan de description** homogène possédant ses unités et règles propres (Lerot 1993 : chap. 4).

Le point de départ sémantique

Les opérations de projection prennent leur départ au niveau assertionnel où sont définies les structures assertionnelles et où le contenu propositionnel est organisé en vue de la communication en fonction de l'intention et de la stratégie communicative du locuteur, du contexte d'énonciation (univers du discours) en vigueur au moment de l'énonciation et des attentes et connaissances qu'il présume chez son interlocuteur.



Les structures assertionnelles sont établies à partir de réseaux sémantiques inspirés de Sowa (1984 ; 1991) et dont la syntaxe est simple. Les réseaux ne contiennent que des nœuds et des liens orientés entre paires de nœuds. Nœuds et liens sont étiquetés.

La théorie sémantique utilisée pour la représentation des contenus propositionnels et des structures assertionnelles comprend trois types d'unités :

(a) les classes génériques, appelées également « concepts » ou « types » ;

(b) les représentations symboliques d'objets individuels (personnes, objets, lieux, temps, événements,...), d'ensembles d'individus ou d'instanciations de classes génériques ;

(c) les fonctions conceptuelles comprenant les liens sémantiques et des constantes conceptuelles comme la cause, la pluralité, la totalité, l'existence, la négation, etc.

Les plans de description

Au plan assertionnel où sont définies les structures assertionnelles (A-structures) succèdent deux plans syntaxiques et le plan phonologique.

Le plan syntaxique le plus profond est celui des F-structures ou structures syntaxiques fonctionnelles.

(a) Il est caractérisé par la présence d'unités lexicales pourvues de leur forme d'expression et de leur catégorie lexicale, par exemple : [NOM : *chat*].

(b) Les unités grammaticales y sont représentées par des formants grammaticaux assortis de leur catégorie grammaticale, par exemple : [NOMBRE : pluriel].

(c) Les unités, qu'elles soient lexicales ou grammaticales, sont organisées dans une structure hiérarchique de constituants.

(d) Les constitués sont définis par leurs constituants et appartiennent chacun à une catégorie syntaxique propre.

(e) Les constructions syntaxiques permettent de définir les fonctions grammaticales profondes des constituants : bases, compléments, adjoints.

Le plan syntaxique le plus proche de l'expression contient les C-structures ou structures en constituants. Les unités y sont des morphèmes organisés en formes de mots et les formes de mots en syntagmes ou en phrases. À la différence des F-structures, l'ordre séquentiel des constituants est pertinent.

Au plan phonologique, la structure (P-structure) est organisée en phonèmes, syllabes, groupes rythmiques et périodes. C'est à ce niveau que le contour intonatoire est défini.

Il est intéressant de remarquer que les formes d'expression typiques de la topicalisation et de la focalisation à savoir les unités grammaticales, les tournures syntaxiques, l'ordre des mots, l'accentuation et l'intonation se répartissent sur ces trois plans de description.

Système de projection

Dans ses grandes lignes, le système de projection consiste à transformer une architecture typiquement dépendancielle en séquences linéaires par l'intermédiaire des structures hiérarchisées en constituants. Afin de garantir les opérations d'unification, nous utilisons des listes dont les éléments sont des paires attribut-valeur (les *facettes*). Le format attribut-valeur a l'avantage de convenir à la fois pour les représentations sémantiques et syntaxiques.

Attributs-valeurs sémantiques :
[isa : *car*], [agent : a], [quant : ALL]
Attributs-valeurs syntaxiques :
[N^0 : *voiture*], [$TEMP^0$: présent], [V^1 : [V^0 : ...] [D^1 : ...]]

Un des premiers problèmes liés à la génération de phrases à partir de représentations sémantiques consiste à déterminer un itinéraire qui permettra de définir des séquences ordonnées d'entités attributs-valeurs.

Représentation des fonctions communicatives

L'illocution

Nous introduisons une instance énonciative (u) associée aux paramètres de l'énonciation : locuteur, allocutaire, temps et lieu de l'énonciation auxquels nous ajoutons :

– le type illocutif ;
– une modalisation énonciative portant sur le type illocutif ;
– le domaine focal portant sur une instance propositionnelle (p) ;
– le topique.

```
u ⌈ agt:   a      "locuteur"                                          ⌉
  | add:   b      "allocutaire"                                       |
  | time:  t      "temps d'énonciation"                               |
  | loc:   l      "lieu d'énonciation"                                |
  | illoc: <type  illoc.> [ mode: <mod. énonciative> ]                |
  | foc.   p      "domaine focal"                                     |
  ⌊ top.   c      "topique"                                           ⌋
```

La prédication

La prédication sera représentée au moyen du symbole (p) désignant une instance de prédication. Celui-ci est associé à trois paramètres : la prédication, l'argument et la modalité éventuelle :

```
p ⌈ modal:  <valeur modale>  ⌉
  | pred:   e  "prédicat"    |
  ⌊ arg:    a  "argument"    ⌋
```

Le domaine focal et le domaine scénique

Le contenu propositionnel est organisé sous forme de réseau sémantique composé d'un certain nombre de connexions orientées. Ce réseau définit un certain nombre de portions cohérentes centrées autour d'un nœud source et définies comme l'ensemble des trajets qui partent de ce nœud source.

Dans le réseau :
w←---x--→y←---z
on peut, à partir de (x), atteindre (w) et (y)
et, à partir de (z), atteindre (y).

Par l'acte d'énonciation, le locuteur fait porter son illocution sur la relation entre un prédicat et son argument. Par l'acte de prédication, le locuteur choisit, parmi les connexions présentes dans le contenu propositionnel, celle qui constituera le **noyau prédicatif**.

Dans le réseau :

w←---x--→y←---z

le locuteur peut faire porter l'illocution sur chacune des connexions suivantes :

x--→w

x--→y

z--→y

Le **domaine focal théorique** est représenté par l'ensemble des trajets qu'on peut atteindre directement à partir du noyau prédicatif tandis que le **domaine scénique théorique** est représenté par les trajets qu'on ne peut atteindre directement à partir du noyau prédicatif.

Dans le réseau :

w←---x--→y←---z

ayant par exemple pour noyau prédicatif : x--→w,

le domaine focal théorique est représenté par : w←---x--→y

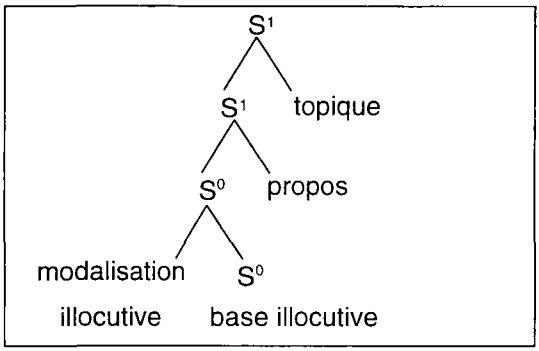
et le domaine scénique théorique par : z--→ y

Cependant, c'est en dernière instance le locuteur qui détermine le contenu à focaliser ou à mettre à l'arrière-plan. Il peut élargir ou restreindre le domaine focal en fonction de sa stratégie communicative. La nature des opérations à effectuer sur les réseaux fait l'objet de recherches en cours.

Projection syntaxique des fonctions communicatives

L'illocution

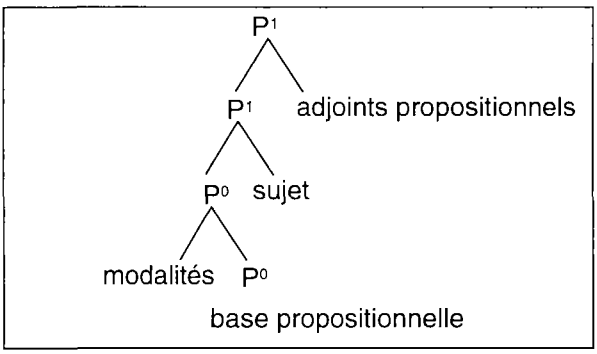
L'illocution constitue la véritable base de la phrase. Elle est présente dans toute phrase-texte ou phrase directrice de texte. Les différents types illocutifs exercent des contraintes spécifiques sur leur complément : une interrogation totale porte sur une proposition, une requête sur une action qui n'a pas encore eu lieu, une ovation et une malédiction sur une personne ou un objet, (par exemple : *Vive la mariée, Malheur aux vaincus !*) tandis qu'une interjection ne demande en principe aucune complémentation (par exemple : *Aïe !*). Pour la représenter nous utiliserons le symbole S^0 désignant la base illocutive et S^1 désignant la phrase formée par S^0 saturé par son complément. Ce que nous avons appelé « propos » (*comment*) se présente comme le complément syntaxique de la base illocutive.



Les modalisateurs illocutifs sont étroitement liés aux bases illocutives : il existe des modalisateurs spécifiques pour les questions, d'autres pour les requêtes, d'autres pour les assertions, etc. Ils sont en outre facultatifs. Ils se présentent donc comme des adjoints de la base (*zero level adjunct*) (Radford 1988 : 255, 261).

Le topique, contrairement à l'illocution, est facultatif. Comme la catégorie du topique est indépendante de la base illocutive, le topique n'est donc pas « régi » par celle-ci et n'est donc pas un complément. Comme une phrase admet plusieurs topiques, la topicalisation est récursive. Si on observe l'ordre des mots dans la phrase, on constate que le topique occupe dans la phrase une position marginale. Ces constatations convergentes nous amènent à considérer le topique comme un adjoint de la phrase, c'est-à-dire comme adjoint de S^1.

La proposition syntaxique



Le propos se présente généralement sous l'aspect d'une « proposition » (au sens syntaxique du terme) qui comprend un « prédicat » (au sens syntaxique du terme) qui en constitue la base, représentée par P^0, et un « sujet » qui se comporte comme son « complément ». En effet, le sujet « sature » la base propositionnelle P^0 et constitue avec cette dernière une proposition pleine, représentée par P^1. En outre, la base propositionnelle a des adjoints qui lui sont propres et qui expriment différentes valeurs modales. Enfin, la proposition pleine est susceptible d'accueillir des adjoints en nombre variable.

D'un point de vue communicatif, le sujet et la base propositionnelle (le prédicat) ne semblent pas liés à une valeur communicative particulière. Les fonctions syntaxiques sont, d'un point de vue communicatif, polyvalentes ou multifonctionnelles (Sasse 1987 : 565). Cependant, on admet généralement que le prédicat syntaxique est le prototype du domaine focal tandis que le sujet est le prototype du topique.

Conclusion

Les hypothèses formulées ici doivent naturellement donner lieu à une vérification par l'étude des expressions topicales et focales dans des langues aux structures diverses.

La topicalisation et la focalisation font partie des problèmes qui ne peuvent trouver leur solution que si on tient à l'esprit le fait que la forme fondamentale du langage humain est l'échange oral face à face et que la priorité de fait accordée à la langue écrite, en jetant le voile sur des phénomènes fondamentaux, rend la description linguistique plus compliquée qu'elle ne l'est réellement.

Le fait que les fonctions grammaticales traditionnelles soient polyvalentes d'un point de vue communicatif explique à la fois le grand intérêt que ces formes présentent pour l'expression écrite et la difficulté de les reconnaître dans des documents écrits.

Enfin, il faut garder à l'esprit que le caractère connu, inconnu, nouveau ou ancien d'une donnée d'information n'est pas pertinent pour l'établissement des fonctions communicatives. Celles-ci sont déterminées par le choix délibéré du locuteur d'activer une portion du savoir de son interlocuteur, d'orienter son esprit vers un contenu déterminé et de susciter un certain type de réaction. La mise en perspective communicative d'un contenu propositionnel n'est rien d'autre qu'une des opérations de la création d'un texte à partir d'un contenu propositionnel donné.

Références

ALTMANN, H. (1981) : *Formen der 'Herausstellung' im Deutschen*, Tübingen.

AUSTIN, J. L. (1962) : *How To Do Things With Words*, Oxford, Clarendon.

BENEŠ, E. (1973) : « Thema-Rhema-Gliederung und Textlinguistik », Sitta, H. et K. Brinker (dir), *Studien zur Texttheorie und zur deutschen Grammatik*, Düsseldorf, pp. 42-62.

CHAFE, W. L. (1975) : « Giveness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View », LI, Ch. N. (dir), *Subject and Topic*, New York, Academic Press, pp. 25-55.

COMBETTES, B. (1983) : *Pour une grammaire textuelle*, Bruxelles, De Boeck-Duculot.

FIRBAS, J. (1964) : « On Defining the Theme in Functional Sentence Analysis », *Travaux linguistiques de Prague*, 1, pp. 267-280.

GADET, F. (1989) : *Le français ordinaire*, Paris, Colin.

GRICE, H. P. (1975) : « Logic and Conversation », Cole, P., Morgan, J., *Speech Acts. Syntax and Semantics*, New York, pp. 41-58.

HAFTKA, B. (1978) : « Bekanntheit und Neuheit als Kriterien für die Anordnung von Satzgliedern », *Deutsch als Fremdsprache*, 15, pp. 157-164.

HAJICOVA, E. et P. SGALL (1987) : « The Ordering Principle », *Journal of Pragmatics*, 11, pp. 435-454.

HALLIDAY, M. A. K. (1967-68) : « Notes on Transitivity and Theme in English », *Journal of Linguistics*, 3-4, Part I (1967), pp. 37-81, Part II (1967), pp. 177-274, Part III (1968), pp. 153-308.

HOCKETT, C. F. (1958) : *A Course in Modern Linguistics*, New York, MacMillan.

JACOBS, J. (1984) : « Funktionale Satzperspektive und Illokutionssemantik », *Linguistische Berichte*, 91, pp. 25-58.

KUNO, S. (1972) : « Functional Sentence Perspective: A Case Study from Japanese and English », *Linguistic Inquiry*, 3, pp. 269-320.

LEROT, J. (1993) : *Précis de linguistique générale*, Paris, Minuit.

LI, Ch. N. (dir) (1975) : *Subject and Topic*, New York, Academic Press.

LYONS, J. (1977) : *Semantics II*, Cambridge, CUP.

RADFORD, A. (1988) : *Transformational Grammar*, Cambridge, CUP.

REINHART, T. (1981) : « Pragmatics and Linguistics: an Analysis of Sentence Topics », *Philosophica*, 27, pp. 53-94.

REIS, M. (1977) : *Präsuppositionen und Syntax*, Tübingen, Narr.

SASSE, H.-J. (1987) : « The Thetic/Categorial Distinction Revisited », *Linguistics*, 25, pp. 511-580.

SEARLE, J. R. (1970) : *Speech Acts. An Essay in the Philosophy of Language*, Cambridge, CUP.

WEIGAND, E. (1979) : « Zum Zusammenhang von Thema/Rhema und Subjekt/Prädikat », *Zeitschrift für germanistische Linguistik*, 7, pp. 167-189.

ZAEFFERER, D. (1981) : « On a Formal Treatment of Illocutionary Forces Indicators », Parret, H., Sbisà, M. et J. Verschueren (dir), *Possibilities and Limitations of Pragmatics*, Amsterdam.

PARTIE II

24

Traductique et traduction humaine : concurrence ou complémentarité ?

Christine DURIEUX

Université Paris III, Paris, France

• *Abstract* •

Initially, research in computational linguistics was aimed at developing automated translation systems. The goal was to substitute machine translation for human translators. The whole effort meant a competition with human or natural translation.

Now, automation tends to materialize as tools that are primarily intended to make machine translation possible but also that are likely to help human translators in their tasks.

In particular, while hypernavigation tools for large documentations are developed as a way to analyse natural texts to be processed by computer, they can also be efficiently used by translators to get knowledge they need to execute technical translations.

Finally, there is a bridge between research in computational linguistics and application to human translation, which leads to conclude in favour of complementarity of both translation forms.

Introduction

Trente milliards de dollars, tel est le coût annuel estimé[1] des traductions effectuées de par le monde. Il n'y a pas lieu de s'étonner qu'une activité de cette envergure ait fait l'objet de travaux en vue de son automatisation.

Dans un premier temps, la démarche s'est inscrite dans la droite ligne de la révolution industrielle, l'objectif étant, à terme, de remplacer l'homme par la machine.

1. Estimation avancée par Jaime Carbonell, Directeur du *Carnegie Mellon Center for Machine Translation* à Pittsburgh, Pennsylvanie.

Les premiers résultats, qui apparaissent comme les toutes premières applications non numériques de l'informatique, ont suscité un réflexe de défense chez les traducteurs, qui se sont sentis menacés dans leur savoir-faire. Certes, la concurrence était ouverte, mais la qualité était loin d'être au rendez-vous. D'ailleurs, les traducteurs n'ont pas manqué de se gausser des absurdités produites par la machine.

Aujourd'hui, le manichéisme n'est plus de mise. Il ne s'agit plus d'être pour ou contre la traductique, de rejeter la traduction machine pour porter aux nues la traduction humaine qui, soit dit en passant, ne le mérite pas toujours, tant s'en faut. Si les traductions qui circulent étaient toujours le fruit d'une véritable démarche de réécriture, la machine n'apparaîtrait pas, tout au moins pas encore, comme un recours possible. Mais il n'est malheureusement pas rare que les traductions ne soient que le piètre aboutissement d'un transcodage – nous en avons tous, sans aucun doute, de multiples exemples en mémoire – auquel cas la machine devient une vraie concurrente : plus rapide et moins chère. Dans les industries manufacturières, quand la machine a fait son apparition, on a d'abord cherché à l'utiliser pour exécuter ce qui était jusque-là fait à la main. Ensuite, on a imaginé ce qu'on pourrait faire d'autre grâce à elle. Il en va de même dans le domaine de la traduction. À l'heure actuelle, la masse des textes qu'il serait utile de traduire est très supérieure à la somme des textes effectivement traduits. La traductique devrait donc permettre de traduire un grand nombre de textes qui, faute de temps et de moyens, restent non traduits. Même si la machine prend un peu du travail du traducteur, c'est une proportion infime, qui d'ailleurs est largement compensable, compte tenu du potentiel énorme du marché.

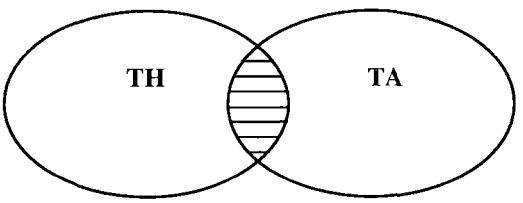


FIGURE 1 : Relation entre traduction humaine et traductique.

Contrairement à la robotique, qui a chassé les hommes des ateliers et des usines, la traductique trouve progressivement sa place à côté de l'activité humaine. Déjà, en pratique, sur le plan de l'exercice de la profession, on constate une complémentarité des deux formes de traduction. De fait, elles visent des finalités différentes. La traductique fournit des textes principalement à usage interne, alors que la traduction humaine reste irremplaçable pour des textes destinés à une large diffusion, à exercer un impact sur le comportement du lecteur : à instruire, à convaincre, à véhiculer une image, notamment. Dans ces circonstances, le mot est porteur d'une valeur de communication. Les tristes listes de synonymes ne sont ici d'aucune utilité, si tant est qu'elles puissent l'être à un moment quelconque de l'opération traduisante effectuée par un traducteur humain. Le connoté l'emporte sur le dénoté. Tout est affaire de cible visée et d'impact recherché.

La finalité est donc un premier critère de tri ; le second étant le volume. En effet,

des documents très volumineux, de préférence répétitifs et à forte densité terminologique, se prêtent plus particulièrement à un traitement par machine. Sur le strict plan du marché de la traduction, il semble donc que l'on puisse déjà conclure à la complémentarité de ces deux formes de traduction.

La traductique

La décomposition du processus de traduction en vue de son exécution par la machine donne cinq étapes : la saisie, l'analyse, le transfert, la synthèse et la sortie. La saisie et la sortie ne sont pas propres à l'activité traduisante ; ce sont des étapes communes à de nombreuses applications, pour lesquelles les possibilités techniques sont aujourd'hui très diverses : saisie au clavier, lecture optique, copie et conversion de fichier, d'une part, affichage à l'écran du document tel qu'il sera imprimé et impression sur des imprimantes de plus en plus performantes, d'autre part.

Ce qui est spécifique à une situation de traduction, au niveau de la saisie, c'est par exemple la possibilité de conserver les attributs d'édition du texte original pour les transférer dans la traduction produite : caractères des titres et des sous-titres, présentation des tableaux, numérotation des listes, gras, italique, souligné, etc. Au niveau de la sortie, c'est notamment l'impression du texte original et de sa traduction en parallèle, et l'insertion de graphiques avec leurs légendes.

Les outils qui permettent de mener à bien ces deux étapes sont très efficaces dans les systèmes de traductique et sont d'une grande utilité dans la traduction humaine, mais on pourrait les qualifier d'outils d'intendance.

Avec l'analyse, on entre dans le vif du sujet. Pour cette étape du processus, des outils ont été développés. On relève à l'heure actuelle deux grandes catégories : les outils fonctionnant sur le principe des statistiques et les outils faisant appel à des connaissances linguistiques. Dans un cas comme dans l'autre, le but visé est de prendre en compte le contexte généralement limité à un microcontexte de quelques mots, pour parvenir à une désambiguïsation lorsque des unités linguistiques du texte original se prêtent à plusieurs interprétations. Cette analyse est lexicale et/ou syntaxique. Les travaux se poursuivent dans ce domaine, et les résultats sont extrêmement intéressants.

Ensuite, vient le transfert, qui consiste à transformer la représentation obtenue à l'issue de la phase d'analyse en une représentation équivalente valable pour la langue d'arrivée. L'algorithme de transfert est propre à chaque langue d'arrivée, puisque les structures de phrase diffèrent considérablement d'une langue à l'autre, même entre nos langues européennes, sans parler des langues dites orientales. Disons-le tout de suite, c'est ce que la machine réalise le moins bien ; c'est le point sur lequel achoppent la plupart des systèmes jusqu'à présent.

Au cours de la phase de synthèse, la représentation construite pour la langue d'arrivée est convertie en langue naturelle. Bien entendu, la qualité du résultat est subordonnée à celle du résultat de la phase précédente, mais il faut bien avoir conscience du fait que les difficultés s'ajoutent alors, ou plutôt se multiplient, rendant le produit final très aléatoire.

La démarche interprétative

L'opération traduisante, telle qu'elle est exécutée par un traducteur, s'articule en deux temps majeurs – la compréhension et la réexpression – chacun étant subdivisable en phases secondaires. Contrairement à ce qui est parfois présenté, ces phases ne sont pas séquentielles mais intimement imbriquées. Par exemple, pour que soit mené à bien le premier temps – la compréhension du texte à traduire – les processus mentaux suivants se déroulent : lecture, mobilisation des connaissances linguistiques et des connaissances thématiques, fusion des significations des unités linguistiques du texte et des éléments pertinents du savoir préalablement acquis, construction d'un sens, déverbalisation, avec de nombreux allers et retours jusqu'à l'appréhension du sens, matière abstraite qui fera ensuite l'objet de la réexpression.

On peut se demander quelle est, dans ce type de démarche, la place des outils informatiques développés pour l'exécution ou l'assistance du processus de traduction par la machine. *A priori,* la réponse est *aucune.* De fait, les résultats fournis par un analyseur lexical ou syntaxique, par exemple, n'offrent guère d'intérêt au traducteur. L'opération d'analyse se fait dans son esprit, inconsciemment, au fil de la lecture, et le fait d'avoir connaissance de la représentation sous forme d'arbre ou de liste d'une portion de texte ne le dispense pas de lire cette même portion de texte original et donc, d'en faire lui-même spontanément l'analyse. Par ailleurs, comme la traduction humaine ne fonctionne pas par décodage-transcodage-recodage, bénéficier des résultats d'une ou de plusieurs de ces opérations parcellaires ne sert pas le traducteur. En effet, il n'y a apparemment pas de recoupement entre le processus de la traduction avec l'intervention de la machine et la démarche interprétative de la traduction humaine. Le premier porte sur des unités linguistiques, même regroupées, et la seconde porte sur le sens que, jusqu'à présent, seul l'esprit humain peut construire ou reconstruire.

Avec l'accélération du progrès technologique et le fantastique accroissement de la vitesse et de la puissance des microprocesseurs et de la capacité des mémoires, l'ordinateur peut battre les plus grands champions d'échecs. En effet, dans une partie, à chaque phase de jeu, le nombre de coups possibles est certes très grand, mais c'est un nombre fini ; c'est pourquoi l'ordinateur peut calculer pour chaque coup jouable toutes les possibilités de riposte et cela jusqu'à l'issue de la partie. Si une langue était un ensemble fini de mots et de règles de syntaxe, l'ordinateur pourrait sans doute réaliser les mêmes prouesses, mais l'expérience prouve qu'il n'en est rien. D'ailleurs, ne qualifie-t-on pas une langue de « vivante » ? Des mots naissent chaque jour, d'autres disparaissent, d'autres encore s'enrichissent d'acceptions nouvelles, d'autres s'appauvrissent, se déforment, puis se reforment. La syntaxe s'assouplit, se déstructure. Sans cesse, de l'inédit jaillit qui déborde de l'analyse soigneusement consignée et échappe à toute démarche systématique.

À ce stade, on pourrait être tenté de faire valoir un constat d'échec, de déclarer que la traductique et la traduction humaine sont irréconciliables et de conclure à une concurrence pure et dure, sans appel, entre ces deux formes de traduction. Ce serait formuler un jugement très primaire, mais ce serait aussi, sur le plan *pratique,* priver le traducteur d'outils de référence et de documentation dont il a le plus grand besoin, et ce serait enfin, sur le plan *théorique*, refuser la réflexion et le débat.

Aspect pratique

Il existe effectivement une passerelle entre recherche en traductique et application traductologique : elle se situe non pas au niveau du transfert mais bien en amont.

Les travaux effectués dans le domaine de l'analyse de la langue, qui visent à répertorier le lexique et ses agencements usuels, ont pour but de faire progresser la traductique, d'affiner les systèmes, de réduire la marge d'erreur de la machine, que les outils ainsi développés soient des analyseurs autonomes ou intégrés à des systèmes complets de traduction (c'est pourquoi le terme de traductique est préféré ici aux sigles TA ou TAO).

Or, les recherches en matière d'analyse sont aussi sous-jacentes au développement d'outils d'hypernavigation, c'est-à-dire de navigation dans des hypertextes. C'est là que se situe la passerelle : l'hypertexte est une source privilégiée de référence et de documentation pour le traducteur et, à ce titre, des travaux menés dans le but premier de le supplanter, en lui substituant un processus antinomique par rapport à sa démarche naturelle, jouent finalement en sa faveur et viennent l'assister dans son effort d'application de la méthode interprétative de la traduction. En réalité, les résultats des recherches en traductique étaient attendus au niveau du transfert entre deux langues et, en fin de compte, ils interviennent le plus utilement en amont de cette opération : au stade de la consultation de sources documentaires et de références.

Pour le traducteur, l'hypertexte offre une richesse exceptionnelle, en lui permettant de naviguer, à son gré, selon sa forme d'esprit et ses besoins ponctuels, à travers une grande masse de documents.

Si le traducteur, désireux de se documenter sur un certain sujet, se reporte à un manuel, par exemple, l'auteur du manuel lui impose sa démarche, le contraint à lire les chapitres les uns après les autres, faute de quoi le lecteur risque de ne plus comprendre la matière exposée. Certes, la présence d'un index peut lui permettre de consulter directement la partie qui l'intéresse sans repasser par tout le cheminement de l'ouvrage, mais cela présuppose chez lui la connaissance préalable de ce qui est dit dans tout ce qui précède le passage consulté, et la mobilisation de ce savoir au moment opportun pour comprendre le passage lu, ce qui revient à lui imposer une démarche mentale.

Déjà, avec une encyclopédie, le lecteur peut s'affranchir d'une lecture linéaire imposée et, par le biais des corrélats, cheminer à sa guise dans la masse des informations offertes. Aucun parcours de lecture ne lui est imposé entre les articles, mais il faut bien reconnaître qu'à l'intérieur d'un article, la lecture reste le plus souvent nécessairement linéaire, la lecture d'une section présupposant que soit acquis le contenu des sections précédentes.

Avec l'hypertexte, rien de tel : les parcours de lecture sont libres ; grâce au multifenêtrage, plusieurs documents peuvent être affichés simultanément pour être lus sur un même espace-écran ; le lecteur dispose de fonctions interactives lui permettant d'opérer des choix et de réorienter son parcours au fil de la lecture. On mesure toute l'utilité d'un tel produit pour le traducteur, surtout si la consultation peut se faire en ligne à partir d'un logiciel de traitement de texte.

Aspect théorique

Texte, contexte, hypertexte : voilà les trois pôles de la réflexion.

Il est vrai que les travaux en traductique ont formidablement progressé dès lors que le simple transcodage de mots a été abandonné au profit de la prise en compte du contexte. Mais quel contexte ?

Pour les chercheurs en informatique linguistique ou en traductique, il s'agit le plus souvent du microcontexte verbal, c'est-à-dire des quelques unités linguistiques qui entourent le mot étudié. Même si l'opération porte sur l'ensemble du texte, ce n'est encore qu'un contexte verbal, auquel cas, l'opération reste au niveau de l'analyse de la langue et ne permet pas d'accéder au sens.

Pour les chercheurs en traductologie, tenants de la théorie interprétative de la traduction, il s'agit bien entendu d'associer au contexte verbal qui désambiguïse les significations, le contexte cognitif qui recouvre les connaissances du sujet et le contexte situationnel qui englobe les paramètres propres aux conditions et circonstances de l'émission du texte. C'est à la lumière de ces trois contextes, dont la fusion se réalise – dans le meilleur des cas – dans l'esprit du traducteur, que ce dernier reconstruit le sens du texte original.

Avec l'avènement de l'hypertexte, de quel contexte y a-t-il lieu de tenir compte ? Le contexte se construit par le jeu de l'établissement de liens, de la constitution d'associations entre éléments de connaissance. À la lecture du texte à traduire, les mots déclenchent toute une série d'associations d'idées, d'images qui se forment à l'esprit, qui se superposent et se télescopent. Finalement, au fur et à mesure de la progression de la lecture, se forme un grand réseau sémantique de type hypertextuel. *Le contexte n'est donc pas une entité objective fixe qui préexiste au texte* ; ce n'est pas un ensemble unique et fini, valable pour un texte et déterminé par celui-ci. Le contexte est au contraire un montage subjectif à géométrie variable, c'est un ensemble évolutif en perpétuelle construction et reconstruction, qui se forme postérieurement à la production du texte, à chaque nouvelle lecture.

Bref, la formation du contexte est une fonction bijective à plusieurs variables qui suit l'expression mathématique suivante :

$$C = f(T, L, t)$$

où : C est le contexte
T est le texte
L est le lecteur
t est le temps

En fin de compte, même s'il semble évident d'affirmer que le sens d'une portion de texte est fonction du contexte, il apparaît de plus en plus pertinent de dire que chaque portion de texte contribue à la formation du contexte en un réseau constamment composé puis recomposé.

Somme toute, peut-être est-ce cette dynamique qui échappe à l'heure actuelle à toute tentative de modélisation.

Conclusion

Même si, sur le plan de la *méthodologie,* on est tenté de parler de concurrence, voire d'antagonisme, entre traductique et traduction humaine, ce constat doit être nuancé. Certes les deux processus ne sont pas conciliables et aboutissent à des produits de nature radicalement différente, promis à des finalités tout aussi différentes.

Toutefois, on observe une complémentarité de ces deux formes de traduction sur le *marché.* En raison de la diversité des exigences – qualité, usage, volume, délai – et des moyens – budget, personnel qualifié, équipement, documentation – traductique et traduction humaine peuvent coexister sans se cannibaliser.

Dans le domaine des *outils,* les efforts consacrés à la traductique se sont soldés notamment par le développement de banques de données textuelles et terminologiques, d'analyseurs lexicaux et syntaxiques, de correcteurs orthographiques et d'outils de construction d'hypertextes et de navigation, qui se révèlent des auxiliaires très précieux pour le traducteur humain.

Enfin, en matière de *recherche*, on observe une fertilisation croisée entre les travaux des chercheurs en traductique et les travaux des chercheurs en traductologie, les uns stimulant la réflexion des autres, et inversement. À l'heure actuelle, la complémentarité est manifeste, comme en témoigne notre présence à tous, ici, pendant ces trois journées.

Références

BALPE, Jean-Pierre (1990) : *Hyperdocuments, Hypertextes, Hypermédias*, Paris, Eyrolles.

GROSS, Maurice (1986) : *Grammaire transformationnelle du français,* Vol. 1 : *Syntaxe du verbe* ; Vol. 2 : *Syntaxe du nom* ; Vol. 3 : *Syntaxe de l'adverbe*, Paris, Asstril.

LEVY, Pierre (1990) : *Les technologies de l'intelligence, l'avenir de la pensée à l'ère informatique*, Paris, La Découverte.

25

La représentation des connaissances en terminologie assistée

Pierre LERAT

Université Paris XIII, Villetaneuse, France

• *Abstract* •

Technical translation deals with conceptual relations, but a translator works with morphological units and syntactic structures: for a human translator knowledge representation is not independent from natural languages. For example, links between concepts like share *(or* stock*),* shareholder *(or* stockholder*),* partner, share capital *(or* capital stock*),* debenture, truncation *(or* dematerialization*),* securitization *(or* securitisation*),* bondholder *(or* bond holder*),* bond, portfolio, company *(or* corporation*) and* security *(or* paper, *or* instrument, *or* warrant, *or* document, *or* certificate*) are both logical and linguistic ones.*

*In a French-English-German terminological data base in the field of contracts (Diké, Université Paris-Nord, thousand concepts), the same word can be both the name of an hyperonym and its hyponym (*en. security 1 = fr. titre, security 2 = fr. valeur mobilière ; fr. obligation 2 = en. bond, obligation non garantie = en. debenture).

There are two kinds of (linguistic) semantic models: componential or relational. One evaluates terminological works using IS A (Meyer et al.*) and PART OF (ISTI), but social sciences also need functional (entity/relation) links. For example,* securitization *is a predicate linked to three entities:* security *(the goal),* loan *(the source) and* company *(the agent).*

From a linguistic point of view, these conceptual connectabilities *(Picht) are not easy to see in a text, because they are neither syntagmatic nor morphological. A terminological data base is very useful if it contains not only syntagmatic collocations and morphological series but also conceptual connectabilities.*

L'exposé que voici tient pour vraies quelques idées qui seront développées dans un livre à paraître (Lerat 1994).

1. En TAO, « le problème n'est pas tant un manque de puissance informatique qu'un niveau insuffisant des connaissances linguistiques » (Carré *et al.* 1991 : 181).

2. Les difficultés de la traduction se situent dans une large mesure au niveau du syntagme, à cause des « collocations non prédictibles » (Kromann 1990 : 24).

Exemple : les valeurs mobilières se « négocient », on opère sur elles des « transactions ».

3. Les difficultés de la traduction technique tiennent aussi à la méconnaissance des liens notionnels, d'où le besoin de prendre en compte des « crochets terminologiques » au sens de R. Dubuc, c'est-à-dire des « appariements de notions » (1992 : 54) ou, dans une approche néo-wüstérienne, des « connectabilités » au sens d'H. Picht (1990), c'est-à-dire des liens conceptuels.

4. Les connaissances sont explicitables sous forme de propositions, c'est-à-dire de jeux plausibles de prédicats et d'arguments.

5. Les connaissances techniques relèvent de « définitions conventionnelles » (Martin 1992 : 68), autrement dit les liens notionnels n'obéissent pas à une rationalité purement linguistique.

Exemple : l'action au sens judiciaire (celle qu'on intente) et l'action au sens boursier (celle qu'on négocie) relèvent de connaissances « domaniales » et non pas « langagières » (voir Melby 1991).

6. En revanche, linguistiquement, les noms des notions sont des substantifs, des verbes, des adjectifs et des adverbes, simples ou composés, analysables morphologiquement et syntaxiquement, et en ce sens leur sémantique est l'interprétation de relations grammaticales prédictibles.

L'hypothèse générale présentée ici est que la représentation des connaissances en vue de la traduction doit à la fois rendre compte des liens notionnels selon une méthodologie non linguistique, tirée de la logique vériconditionnelle, et être évaluable linguistiquement, c'est-à-dire contrôlée morphologiquement et syntaxiquement. Cette approche exclut un niveau de représentation interlinguistique, mais se prête à un codage numérique des notions, qu'elles soient partagées par plusieurs langues ou propres à une culture scientifique ou technique particulière.

Exemple : si l'anglais *capital stock* peut traduire à la fois en français *capital social* et *capital actions*, on prendra en compte cette réalité en distinguant trois notions numérotées différemment.

Le champ notionnel pris comme exemple ici est celui des valeurs mobilières. Les modèles de représentation des connaissances évoqués sont ceux de la sémantique linguistique.

Le champ notionnel considéré : les valeurs mobilières

Les termes pris en compte sont ceux qui concernent les valeurs mobilières dans une

base de données consacrée à la terminologie des contrats. Cette base, Diké, du nom de la justice en grec ancien, a été réalisée par l'Université de Paris-Nord en collaboration avec l'Université juridique de Paris II et le Service de terminologie de la Cour de Justice des Communautés européennes, avec le soutien du ministère de la Recherche. Elle traite un millier de notions en français, avec des équivalents en anglais et en allemand, et le logiciel utilisé est MERCURY-TERMEX.

La liste des termes retenus ici à titre d'illustration est la suivante : *action, actionnaire, associé, capital social, dématérialisation, obligataire, obligation, portefeuille, société, titre, titrisation, valeur mobilière.* Il s'agit de notions assez largement partagées, mais l'idéal terminologique de biunivocité est contrarié par une double variabilité, linguistique et juridique.

La variabilité linguistique s'observe dans des concurrences bien établies. Elle est courante pour les noms de notions nouvelles : par exemple *titrisation* est usuel en Europe, mais Vincent retient *titralisation* et atteste *sécuritisation,* et le Crédit suisse propose *mobiliérisation* pour traduire *securitization* (USA) et *securitisation* (GB, selon TERMIUM).

La variabilité juridique pose des problèmes de gestion autrement délicats dans une base de données. Pour écarter le trop complexe (juridiquement), tenons-nous-en à des notions d'expérience courante et voyons comment traiter leur sémantique, puisque c'est de méthodologie qu'il s'agit.

Évaluation de modèles sémantiques

Les deux options majeures sont soit un modèle componentiel soit un modèle relationnel. Le premier, où les unités de compte sont non pas des termes du vocabulaire considéré mais des noms métalinguistiques de propriétés sémantiques telles que *animé* ou *matériel,* est malaisément falsifiable car, comme le dit N. Ménard, « comment faire le décompte d'unités aussi peu tangibles que les sèmes, les noèmes, les traits ou les primitifs sémantiques ? » (1989 : 158). La terminologie des caractéristiques, au sens de Wüster, postule que les traits sont des propriétés des objets, mais une ontologie juridique tombe moins sous le sens qu'une ontologie des moteurs à explosion, par exemple.

Les modèles relationnels présentent plus d'intérêt pour la terminologie, si l'on en juge par les options qui ont fait l'objet de présentations dans les revues spécialisées ces dernières années. Les arborescences construites sur le modèle de l'hyperonymie ont montré leur intérêt en documentation, en I.A. et en terminologie (voir Lerat 1990), et aussi leurs limites. Un prédicat très abstrait comme SORTE DE a beaucoup d'intérêt pour classer des types d'objets tels que des imprimantes ou des supports électroniques (voir Meyer *et al.* 1992) ; dans le cas des valeurs mobilières, de petites arborescences sont également concevables, mais un hyperonyme comme *ensemble,* pour *portefeuille*, est trop général pour être utile, et la dématérialisation est une *opération*, autrement dit SORTE DE ne conduit ici à rien de suffisamment technique pour un réseau notionnel.

L'Institut supérieur de traducteurs et interprètes de Bruxelles mise davantage sur

les liens partitifs, qui manifestent effectivement un rendement appréciable dans les travaux de P. Merten (1992) et de M. Van Campenhoudt (voir Blampain *et al.* 1991). La situation est moins favorable ici : une valeur mobilière est une partie d'un portefeuille, mais seulement à partir de la deuxième ; ou encore, juridiquement, on ne saurait dire que l'associé, personne physique, est une partie de la société, personne morale.

Sur le modèle des facettes en documentation, on peut aussi recourir à des prédicats *ad hoc*, en fonction de la thématique considérée. C'est ce que fait I. Meyer avec DEGREE OF WRITABILITY (1991 : 234), et aussi P. Merten dans l'exemple suivant : « L'angle des textures EST LA REPRÉSENTATION DE texture » (1992 : 217). Au titre des propriétés saillantes pour les experts, un candidat possible dans le cas présent serait DÉTENTEUR (exemple : un obligataire EST DÉTENTEUR DE obligations, une obligation A POUR DÉTENTEUR un obligataire). En fait, la nature des choses juridiques fait que l'obligataire peut être soit propriétaire, soit détenteur, donc le prédicat DÉTENTEUR n'est ni nécessaire ni suffisant.

D'où le parti, dans Diké, de renoncer à des prédicats substantiels, du type des facettes, au bénéfice d'une structuration pauvre mais universelle : le modèle régissant/dépendant, dont voici les principes. C'est un modèle fonctionnel, au sens où l'on parle de *fonctions syntaxiques* ; ainsi, *titrisation* s'applique à *société,* (comme agent) et à *prêt* (comme objet), puisque par la titrisation « une société convertit des prêts en titres négociables » (Vincent 1988). On pourrait bien entendu, en dénominalisant, utiliser une grammaire des cas, mais il est plus économique et moins aléatoire de prendre en compte seulement des possibilités de groupes nominaux tels que *titrisation par* x (si *x* est une dénomination générique des sociétés) et *titrisation de* y (si *y* est une dénomination générique des créances). On n'a pas non plus à préjuger du caractère plus ou moins concret de ces entités (*prêt* est lui-même un nom prédicatif en tant que nom d'action et un nom concret comme dénomination de ce qui est prêté). Un autre avantage est que, comme la « connexion sémantique » chez Tesnière, ce lien syntaxique universel n'est pas lié aux parties du discours : le substantif peut être régi par un verbe ou par un substantif, l'adverbe est dépendant d'un verbe, d'un adjectif ou d'un adverbe, il ne s'agit pas de valences verbales ni de cas nominaux mais de détermination, de rection ou de *gouvernement*, selon les phénomènes et selon les écoles ou, à l'inverse, de dépendance, de régime ou de complément en général. Cette généralité rappelle à première vue les fonctions lexicales de Mel'čuk, mais elle ne préjuge pas de la pertinence de primitifs sémantiques là où il s'agit de connaissances spécialisées.

On imagine mal un réseau notionnel, même intralinguistique, où seraient gérables toutes les opérations que peut réaliser un sujet de droit, ou même une personne morale, ou même tel type de personne morale. En revanche, entre l'utopie d'une mise à plat des connaissances spécialisées et la misère des grandes banques de données terminologiques en matière de liens notionnels utiles, le modèle régissant/dépendant permet de rendre compte de liens présentant l'intérêt d'être des collocations terminologiques. Ainsi, *négocier dix Total* est analysable par effacement à partir de *négocier dix actions (de) Total,* où *Total* est un nom de société et où *négocier* constitue un régissant approprié pour *action*. Il suffit de pouvoir utiliser à la fois, dans une base de données relationnelle, des connectabilités propres au vocabulaire de la bourse (comme *négocier - action*) et une liste de données factuelles propres à cet univers de connaissances (comme la série des noms de sociétés cotées en bourse auxquelles on a décidé de s'intéresser).

La validation des relations de sens

Le continuum entre unitermes et pluritermes, qui rend aléatoires les nomenclatures en terminologie, résulte de la variation linguistique (condensation et paraphrase, emprunt et calque etc.). En outre, les expressions ne sont des termes que pour autant qu'elles dénomment de façon conventionnelle des connaissances techniques partagées. Autant les analyses de cooccurrents peuvent être utiles pour inventorier un champ notionnel si le corpus est technique et homogène, comme c'est le cas pour *coter, émettre* ou *négocier* à propos d'*action* dans l'ouvrage de B. Cohen (1986), ou encore quand A. Kukulska-Hulme dépouille une documentation pour usagers des tableurs (1989), autant les inventaires de collocations tout venant courent le risque de manquer de pertinence, plus exactement de « connectabilité au niveau conceptuel » (Picht 1990 : 37).

De même que la collocation terminologique n'est pas prédictible à partir des seules règles syntaxiques générales, la connectabilité varie selon les domaines de connaissances : rien qu'en droit, *titre* est associable à *noblesse* ou à *métal précieux* aussi bien qu'à *portefeuille*, et l'on ne peut rien tirer d'un tel chaos lexical. Il en va de même pour la morphologie : *titre* fait partie de la famille de *titrage* et de *titulaire*, pour ne rien dire des sens non juridiques, mais un commencement de clarification n'apparaît qu'à partir du moment où *titrage* est en relation avec *métal précieux* dans la limite d'une valeur particulière de *titre*, et où l'on a résisté à la tentation de faire comme si le titulaire était le détenteur de titres, alors que c'est le détenteur *en titre* (voir le *Vocabulaire juridique,* 1992).

Si le réel des langues spécialisées est une dimension au moins morphologique, syntagmatique et fonctionnelle, la TAO a besoin d'une triple analyse linguistique des termes. On le sait, le traducteur doit communiquer des informations, le traducteur technique des informations techniques, et il doit le faire au moyen d'énoncés syntaxiquement plausibles et morphologiquement corrects. Il y a donc lieu de lui fournir des instruments de contrôle indépendants les uns des autres permettant non plus des *équivalents partiels* sans mode d'emploi mais des règles terminologiques telles que les suivantes :

1. (règle syntagmatique juridique) : « si *action de société,* alors *action* = *en. share* = *de. Aktie* etc. »
2. (règle morphologique juridique) : « si *actionnaire,* alors *action 2* » ; si *actionner* tr., alors *action 1* (action en justice)
3. (règle pragmatique juridique) : « si *portefeuille,* alors *action 2* ».

Liste des liens notionnels utilisés dans Diké

g. = « a pour genre prochain »
p. = « est une partie de »
d. = « dépendant de »
r. = « régissant de »
s. = « synonyme de »
a. = « antonyme de »

Le champ notionnel des valeurs mobilières en français

N.B. : * veut dire « au sens défini dans la base »
action 2 : g. valeur mobilière*, p. capital social*, r. associé*, actionnaire*, société commerciale*, s. action de société, a. obligation 2*
actionnaire : g. associé*, d. action 2*, r. société commerciale*
associé : g. membre, d. société*
capital social : r. société commerciale*
dématérialisation : r. titre*
obligataire : g. créancier*, d. obligation 2*
*obligation 2** : g. titre*, p. emprunt, r, obligataire*, a. action 2*
portefeuille : r. valeur mobilière*
société commerciale : g. société*
titre : g. certificat
titrisation : r. société commerciale*, prêt*
valeur mobilière : g. titre*, d. portefeuille*

Références

BLAMPAIN, D., PETRUSSA, P. et M. VAN CAMPENHOUDT (1992) : « À la recherche d'écosystèmes terminologiques », A. Clas et H. Safar (dir), *L'environnement traductionnel. La station de travail du traducteur de l'an 2001,* Actes du Colloque de Mons (Belgique), actualité scientifique, Sillery, AUPELF-UREF, Presses de l'Université du Québec, pp. 255-271.

CHABRIDON, J. et P. LERAT (1993) : « Terme et famille de termes », *La Banque des mots,* 5, pp. 55-63.

CARRE, R. *et al.* (1991) : *Langage humain et machine,* Paris, Presses du CNRS.

COHEN, B. (1986) : *Lexique de cooccurrents. Bourse, conjoncture économique,* Montréal, Linguatech.

DIKÉ, base de données consacrée à la terminologie des contrats en français, avec des équivalents en anglais et en allemand, Université Paris-Nord.

DUBUC, R. (1992) : *Manuel pratique de terminologie,* 3e éd., Brossard, Linguatech.

EURODICAUTOM, banque de données terminologiques de la Commission des Communautés européennes, disque optique compact TermDoko.

KROMANN, H. P. (1990) : « Selection and Presentation of Translational Equivalents in Monofunctional and Bifunctional Dictionaries », *Cahiers de lexicologie,* vol. LVI, 1-2, pp. 17-26.

KUKULSKA-HULME, A. (1989) : « Dictionnaires informatisés et traduction », *Meta,* vol. 34-3, pp. 533-538.

LERAT, P. (1990) : « L'hyperonymie dans la structuration des terminologies », *Langages,* 98, pp. 79-86.

LERAT, P. (à paraître) : *Les langues spécialisées,* Paris, PUF.

Lexique boursier (1992) : Zurich, Crédit suisse.

MARTIN, R. (1992) : *Pour une logique du sens*, 2e éd., Paris, PUF.

MELBY, Alan K. (1991) : « Des causes et des effets de l'asymétrie partielle des réseaux sémantiques liés aux langues naturelles », *Cahiers de lexicologie*, n° 58-I, pp. 6-43.

MÉNARD, N. (1989) : « Mesure des relations lexico-sémantiques dans des textes scientifiques : problèmes méthodologiques », *Meta*, 34-3, pp. 468-478.

MERTEN, P. (1992) : « Apport des relations notionnelles à la description terminologique », *TAMA '92*, Vienne, TermNet, pp. 201-228.

MEYER, I., BOWKER, L. et K. ECK (1992) : « COGNITERM : an Experiment in Building a Terminological Knowledge Base », *EURALEX '92*, Tempere, *Studia translatologica*, ser. A, vol. 2, pp. 159-172.

PICHT, H. (1990) : « LSP Phraseology from the Terminological Point of View », *Terminology Science and Research*, vol. I, n° 1-2, Vienne, TermNet, pp. 33-48.

TERMIUM, banque de données terminologiques du Secrétariat d'État du Canada sur disque optique compact.

VILLERS, M.-É. DE (1979) : *Vocabulaire des imprimés administratifs*, Montréal, Office de la langue française.

VINCENT, L. (1988) : *Vocabulaire bancaire*, Ottawa, Centre d'édition du gouvernement du Canada.

CORNU, G. (dir) (1992) : *Vocabulaire juridique*, 3e éd., Paris, PUF.

26

Termes et symboles discours hétérogènes. Quelques hypothèses sémiologiques

Yves GENTILHOMME

Université de Franche-Comté, Besançon, France

La terminologie considérée comme science ou discipline est une variante de la lexicologie.
D. Gouadec[1]

• *Abstract* •

The co-presence of terms and symbols in technical-scientific discourse and texts indicates both formal and semantic convergences, divergences and interferences in their syntagmatic and paradigmatic functioning.

This paper draws attention to certain semiological facts and puts forward some hypotheses taking these facts into account. After a short overview of the formal and neological reciprocal influences, the questions of symbolisation criteria plus terminological and symbolic deceptive cognates are analysed.

Several hypotheses are put forward concerning the signified and the existence of morpho-semantic inter-influences while at the same time taking into account the opposition between the future lexicological activity of the linguist and the inventive activity of the research worker on the splitting of the signified into its notional and conceptual aspects, weak and strong arbitrariness and the performance of the defining act... an hypothesis which could lead to the need for rethinking the notions of synonymy, homonymy and polysemy (as opposed to multisemy) with a different impact on common words, terms and symbols.

Finally, the vast heuristic field of the extension to terms and symbols of the lexical functions of the "textless" model is opened up.

1. Daniel Gouadec, 1990, *Terminologie*, p. 14.

Avant-propos

Selon une opinion largement répandue, l'objet d'une terminologie consiste à associer biunivoquement un terme (mot ou groupe de mots[2]) à tout référent (objet concret ou objet de pensée) relevant d'un champ disciplinaire donné, qui permet au spécialiste, s'adressant à un confrère, de le désigner sans équivoque, tout en respectant certains impératifs de cohérence et de gestion[3].

En fait, dans la pratique terminologique, un tel objet – préconisé notamment par Lavoisier[4] – apparaît comme un **projet idéal**, constamment dépassé[5], – à la limite même, pour certains, comme une utopie[6] – pour de multiples raisons dont nous n'aborderons que quelques-unes. Sans condamner sans appel la règle drastique de monosémie, disons qu'il s'agit tout au plus d'une approximation, utile dans une **prime approche naïve**, mais qu'il convient, pour le moins, de nuancer[7].

Il est trivial de dire qu'un terme est tributaire au moins de deux systèmes[8]. D'une part, il doit répondre aux besoins de la discipline (scientifique ou technique) dont il relève ; d'autre part, assumant la fonction de composant communicationnel, il intéresse également la linguistique, notamment la lexicologie[9], ou, mieux, la sémiologie, définie par Saussure comme la science des signes[10].

Limitation du champ d'investigation

Rappelons, à ce propos, l'existence de multiples types terminologiques – plus ou moins normalisés et systématisés selon des consignes (codifiées ou d'usage) spécifiques. Ainsi, pour fixer les idées, les terminologies de la géométrie euclidienne élémentaire, de la mycologie, de l'informatique, de l'industrie alimentaire, du droit civil, etc., aux diverses étapes de leurs développements respectifs, ainsi qu'aux circonstances variées de leur mise en œuvre, ne posent pas des problèmes identiques aux spécialistes, aux apprenants et aux enseignants concernés.

---

2. Si la nécessité de considérer comme des unités terminologiques indivisibles certains groupes spécifiques de mots (*dérivée partielle, courbure totale, variable aléatoire, courbe de Lissajous, hydrogène sulfuré, oxyde puce de plomb...*), extraits d'un contexte technique, relève de nos jours de l'évidence, tel n'a pas toujours été le cas. Notons qu'André Phal, dès 1964, a insisté dans un article : « Les groupes de mots et les problèmes qu'ils posent dans la préédition de textes destinés à l'analyse mécanographique » (*Cahiers de Lexicologie*, vol. IV,1), sur l'importance de ce problème. (pp. 45-60).

Pour une étude générale récente sur les diverses façons de créer un néologisme en langue naturelle, consulter l'ouvrage concis et clair de Jean Tournier, 1991.

3. Voir Gouadec, *op. cit.*, expose cette problématique de façon systématique.

4. Ce projet et son évolution à nos jours ont été étudiés dans François Dagognet, 1982.

5. Pour une approche linguistique plus générale, voir Henri Mitterand, *Les mots français*, (8e éd. 1992) et Rostislav Kocourek, *La langue française de la technique et de la science*, 1982 (2e éd.), ouvrage très documenté, synthèse et mise au point de nombreux travaux antérieurs sur ces questions.

6. D. Gouadec, *op. cit.*, p. 14 : « L'utopie terminologique n'est autre que le vieux rêve des langages référentiels dans lesquels (*i*) une désignation et elle seule correspond à tel objet ou tel concept ou processus ou événement, et (*ii*) réciproquement tel objet ou tel concept ou processus ou événement, et lui seul, ne peut avoir que telle désignation linguistique ».

7. Gentilhomme, 1984, « Les faces cachées du discours scientifique ».

8. On trouvera de nombreuses précisions dans Gentilhomme 1968, 1992, 1993, précisions que nous ne pouvons reprendre dans le cadre de cet article.

9. R. Galisson 1983, *Des mots pour communiquer. Éléments de lexicométhodologie.*

10. Ferdinand de Saussure, 1915, *Cours de linguistique générale.* Pour une synthèses des théories sur cette question : Umberto Eco, *Le signe.*

Ni l'usage, ni la tradition, ni la normalisation, ni la néologie, ni l'attitude psychique consciente ou inconsciente (notamment les non-dits, les allusions discrètes) ne se manifestent de la même façon, pour peu qu'on les observe avec quelque attention.

Que penser des nouvelles interdisciplines empruntant des termes à gauche et à droite, à hue et à dia, tout en poursuivant leur propre cheminement, avec des conceptions idoines de contenu et de rigueur, qui font parfois sursauter les dépositaires des terminologies « pillées » ?

Afin d'éviter les lieux communs ou les généralisations téméraires, nous **limitons notre objet** d'étude en portant notre attention principalement sur les terminologies concernant les quelques matières que nous avons eu l'occasion d'aborder, surtout en situation d'enseignement-apprentissage : soit de la matière proprement dite, soit d'une langue étrangère à des chercheurs et ingénieurs français[11]. Il n'est pas exclu que nos conclusions aient une portée plus large. Il ne nous appartient pas d'en décider.

Approche sémiologique

> Fixons pour chaque H_i un $u_i \in S$ avec $u_i \in (H_i)^{\perp}$.
> D'après 9.12.4 : £ $(P_{\xi}(\text{Face}_i\, P)) = |(\xi|u_i)|$ £$_{Hi}$ $(\text{Face}_i\, P)$,
> donc en vertu de [...]
>
> Extrait de Géométrie, Marcel Berger, 1990, p. 91.

Plus restrictivement encore, nous ne nous intéresserons qu'aux **discours hétérogènes**, écrits ou oraux, faisant un **usage systématique de symboles**, comme c'est le cas, de plus en plus, en mathématiques (voir exergue), logique, physique, chimie.

> Le mot *symbole* peut être interprété de diverses façons. *Le Robert* propose, entre autres, la définition très souple : « Ce qui en vertu d'une convention arbitraire correspond à une chose ou à une opération ». Nous restreindrons ce contenu par l'usage qui en est fait dans les disciplines citées, dont certains aspects seront précisés ultérieurement.

Dans la mesure où le spécialiste veut être précis, rigoureux et concis, il est généralement admis que le fait d'élaborer des démonstrations et d'explorer au-delà du niveau du concret préhensible[12] a pour effet la non-réductibilité de la communication à l'emploi exclusif de mots ou de locutions plus ou moins figées. Les symboles et les schémas y sont porteurs d'information essentielle, difficilement traduisible en un vocabulaire « aseptisé », ressortissant à la seule lexicologie traditionnelle, même enrichie de la terminologie relevant d'un savoir encyclopédique ; d'où l'obligation de dépasser l'approche strictement linguistique par une **approche sémiologique**, prenant en compte symboles et schémas.

---

11. Compte tenu de notre formation et des nécessités nous avons enseigné (niveau secondaire) : les mathématiques, l'optique géométrique, la mécanique rationnelle, l'électricité, la chimie minérale ; puis le russe, notamment au CNRS, dans le cadre de l'Enseignement Préparatoire aux Techniques de la Recherche et dans des Facultés de sciences, à des scientifiques et ingénieurs de haut niveau, ce qui a contribué à élargir l'éventail de notre curiosité ; enfin, en Faculté des lettres : la linguistique générale et la lexicologie. D'où notre intérêt pour les langages scientifiques. Voir Gentilhomme, 1964, *Manuel de russe à l'usage des scientifiques*.

12. Le cas est patent en mathématiques lorsqu'on quitte l'espace tridimensionnel euclidien, ou en physique mathématique, notamment en mécanique quantique où les métaphores rappelant notre univers quotidien sont pour le moins suspectes.

Certes, les idées sous-jacentes au discours technoscientifique, comme le montrent certains ouvrages de vulgarisation ou d'épistémologie, sont exprimables en un discours à tenue littéraire[13]. Sans doute est-il important qu'il en soit ainsi. Bien plus, nous pensons qu'un authentique savant, comprenant en profondeur la problématique qu'il traite, doit être capable d'en rédiger le substrat en un discours simple et clair. Mais c'est là un tout autre problème, en marge du présent champ d'investigation.

Canaux de communication

Que l'on songe seulement à certaines longues formules mathématiques ou physiques, aux matrices à un grand nombre de lignes et de colonnes, aux formules développées complexes de chimie organique !

Même si ces longues chaînes de symboles sont dicibles en langue naturelle, le message-calque interprétant qui en résulte s'avère pénible à percevoir à la fois dans le détail et dans l'entièreté, la compréhension en profondeur exigeant les deux perceptions quasi-simultanées[14].

Insistons sur le fait sémiologique que le spécialiste, au cours d'une conférence, fait appel, en général, à deux canaux communicationnels parallèles – linguistique et symbolique, à la fois partiellement redondants et complémentaires.

Pour mieux faire percevoir la complémentarité des canaux, proposons la métaphore électrique du réseau « monté en dérivation », l'intensité du courant principal étant la somme des intensités des courants dérivés, pondérés en fonction des résistances des dérivations.

Le parler et l'écrire

Tout en parlant, l'orateur écrit des formules, ou montre des formules préécrites, dont il ne nomme pas nécessairement un à un tous les éléments constituants. Il s'attache particulièrement, au fil de la parole, à certains repères ou énonce des raccourcis globalisants, destinés tour à tour à assurer une perception correcte, à focaliser l'attention sur des points particuliers ou à faire saisir l'entièreté d'une suite plus ou moins étendue de symboles.

Pour fixer les idées, supposons qu'un orateur présente la formule (tirée du dictionnaire de physique de J.-P. Sarmant, 1978), correspondant au moment magnétique :

$$M = {}^{1}/_{2} \iiint r \, A \, d\tau$$

Sans doute peut-il la lire formellement, in extenso, « mot à mot » :

[emegal œ̃demi sɔmtriplə dyprɔdɥivektɔriɛl ɛrdete],

Cependant, comme elle peut être vue (écrite ou projetée) par l'auditoire, il peut s'en dispenser et se contenter de la montrer d'un geste en la désignant sans équivoque par sa forme : cette [ɛtegraltriplə]... ou, encore, en fonction de la situation d'énonciation, il peut éviter la lecture formelle et la nommer par son contenu théorique : le [mɔmɑ̃ma ŋetik]...

13. Citons pour l'exemple des noms prestigieux comme : le « dernier homme universel »: Henri Poincaré, des physiciens de haut niveau : Louis de Broglie, Albert Einstein et parmi nos contemporains, l'astrophysicien-poète Hubert Reeves.

14. La préhension aisée des formules exige une longue formation préalable. Il en résulte un effet psychologique pervers. Certains lecteurs se sentent comme paralysés, rien qu'à la vue des formules, alors que d'autres, au contraire, éprouvent une sorte de soulagement de ne pas avoir à parcourir un long texte fournissant la même information.

Ainsi, dans un **système langagier mixte**, les deux sous-systèmes, terminologique et symbolique, interagissent en s'épaulant et en se complétant. De cette interaction résulte l'information complète communiquée.

Ajoutons, à ce propos, que nombre de locuteurs – non-spécialistes s'entend – ignorent non seulement comment prononcer certains symboles, mais, de plus, quels ingrédients linguistiques introduire à ce propos, selon la situation de communication considérée (mots outils, vocabulaire fondamental, vocabulaire général d'orientation scientifique), ce qui pose problème aux enseignants de langue de spécialité[15].

Même dans des cas tout à fait élémentaires, comme :
$2{,}7\ a(bx^2 - cx^3)^4 < y \leq 3{,}01\ a(bx^2 + cx^3)$; $T \approx 2\pi\sqrt{l/g}$; $(m_p) = m!/p!(m-p)!$
le non initié peut hésiter sur les prépositions, les articles à utiliser et sur les pauses appropriées à respecter pour bien faire entendre l'interprétation à saisir :
ab + c se lisant : [abe plysce] et a(b+c) : [a beplysce]

À cela s'ajoute le problème du niveau de langage convenu : Sh*a* s'écrit en toutes lettres *sinus hyperbolique de alpha,* mais se prononce, selon l'état d'esprit du locuteur, de façon plus ou moins académique ou précipitée :
[sinysipɛ rbɔlikdəalfa] ... [ɛ sa ʃ alfa]
Que dire des exemples plus délicats (Sarmant 1978, *ibid.*)
$\overline{\nabla} = U_x \partial/\partial x + U_y \partial/\partial y + U_z \partial/\partial z$ (∇ se lit [nabla])
$\nabla \Lambda (\nabla \Lambda a) = \overline{\nabla} (\nabla \wedge a) - \nabla^2 a$
ou encore d'une de ces longues formules développées de chimie organique, agrémentées de cycles hexagonaux ?

Influence réciproque de surface

Avant d'aborder le problème central des influences réciproques sémantiques entre termes et symboles, rappelons brièvement quelques faits de morphologie de surface bien connus.

Termes générateurs de symboles

Il est évident que nombre de symboles ne sont, à l'origine, que des abréviations plus ou moins réduites de termes (sigles, acronymes), consacrées par l'usage ou en passe de l'être, admettant des variantes diachroniques, géographiques, d'appartenance à une communauté, d'écriture manuelle plus ou moins soignée, de contraintes typographiques, voire de goûts personnels ou relevant de l'humeur de l'instant.

À titre d'illustration, citons :

tangente réduit à tang, tg, tan (actuel) ; *argument cotangente hyperbolique* → argcoth, Arc coth ; *logarithme (quelconque, népérien, vulgaire ou binaire)* → log, $\log_n$, Log, Ln, ln, $\log_{10}$, lg, $\log_2$, Lb ; *exponentielle* (base e) *de* x → exp x, e^x ; *limite de* x → lim x ; *di-*

15. Voir à ce propos Georges Gougenheim et coll. 1958 et 1964, André Phal, 1971, *Vocabulaire général d'orientation scientifique,* la critique du *Français fondamental* par Robert Galisson, 1970 et 1983 : si l'on s'en tient aux vocabulaires ci-dessus, il est impossible de dire quoi que ce soit d'intéressant, et de citer en exemple « la vache fondamentale » qui est dans l'impossibilté de se conduire comme une vache normale : qui ne peut pas brouter dans un pré, qu'on ne peut pas traire, etc., qui est réduite à se promener sur un trottoir. Il en est de même pour les textes technoscientifiques. Il s'ensuit que chaque spécialiste doit se constituer son propre vocabulaire.

vergence (d'un champ vectoriel → V) → div → V ; *gradient* (d'un champ scalaire U) → grad → U...[16]

Critères de « symbolisation »

À l'instar de la question : dans quelles conditions un mot peut-il être considéré comme un terme, demandons-nous : quand une abréviation devient-elle un symbole ?

S'il paraît naturel de considérer comme une abréviation et non comme un symbole, le célèbre « C.Q.F.D. » (c'est ce qu'il fallait démontrer) qui naguère ponctuait les démonstrations des théorèmes, que penser de P.G.C.D. (plus grand commun diviseur) et de P.P.C.M. (plus grand commun multiple), de PH (paraboloïde hyperbolique), de pH (potentiel-hydrogène des chimistes) ?

On peut d'abord distinguer les critères proprement formels des critères syntaxiques.

Souvent la symbolisation se manifeste par l'adoption d'un type particulier de caractères (dans une impression soignée) : gras, italiques, cursifs, ronds, grecs, gothiques, hébraïques... voire déformés, renversés.

> N (aleph, utilisé depuis Cantor pour noter des cardinaux infinis), $\Re$ (ensemble des réels), $\mathcal{P}$ (ensemble des parties d'un ensemble), $\exists$ (quantificateur existentiel), $\forall$ (quantificateur universel), $\in$ et $\notin$ (appartenance et non appartenance) issus de la lettre grecque ε...
>
> D'ailleurs, une différence de type peut être utilisée pour distinguer des entités différentes. Une lettre grasse peut équivaloir à une lettre fléchée.
>
> Certes, l'exploitation du type existe également en langue, mais pas de la même façon. Ainsi, l'italique peut indiquer un emploi autonyme (le mot se désigne lui-même : « La force est la dérivée du moment par rapport au temps » et « Le mot *force* s'écrit avec cinq lettres ») ou une citation.

Dans un cadre de linguistique contrastive, la symbolisation d'une abréviation de vient manifeste lorsque, dans un texte imprimé en caractères non latins (cyrilliques, arabes, chinois...), on conserve les graphies d'origine : lim, log, div, rot... non transcrites, alors que d'autres expressions le sont.

Certains symboles non littéraux se présentent comme des icônes ; de même que, dans le code de la route, **Z** prévient qu'il y a un tournant ou une tête de mort sur un flacon avertit qu'il contient un poison :

⊥ (perpendiculaire), // (parallèle), Δ (triangle), ∠ (angle), ⊃ (contient, en math., implique en logique), ⇒ (implique), $\not\subset$ (n'est pas contenu), etc.

> L'utilisation de tels symboles à l'intérieur d'un texte est parfois jugée sévèrement par des puristes, taxée de laxisme intolérable, laxisme que le temps finit parfois par valider[17].

16. On trouvera de nombreux exemples dans Le Lionnais 1979, p. 791 et sq. ; dans Berger 1990, p. 537 et sq., ainsi que dans Quaterner et Trotignon, 1981, p. 10 et sq., non sans quelques divergences de graphie.
17. Dans son article : *Sur les mots et les symboles*, le logicien Daniel Lacombe,1964, critique l'usage des symboles non conforme à la tradition.

Il est important de souligner l'inventivité néologique luxuriante et l'extrême diversité des symboles non-littéraux. Les flèches, notamment, prennent des formes originales en fonction du rôle théorique qu'elles assument. Pour s'en convaincre, il suffit de jeter un coup d'œil, ne serait-ce que sur les premiers manuels des mathématiques dites modernes.

Le temps finit par normaliser les graphismes qui, à la longue, finissent par acquérir un statut international.

Pour distinguer un symbole d'une simple abréviation, en s'inspirant de la définition du mot comme unité de fonctionnement (voir, entre autres, Lucien Tesnière (1959) ; Bernard Pottier (1974)), on peut proposer le critère syntaxique suivant : un symbole est susceptible de rentrer dans des formules soumises à certaines règles propres à la discipline considérée.

Il convient de rappeler à ce propos qu'il y a formule et formule ! Nous opposons les formules de type **logico-mathématique** aux **formules-images**.

Si les premières sont astreintes à des règles explicites de bonne formation, voire de dérivation et peuvent être déclarées : justes, fausses ou dénuées de sens (mal formées) ; les secondes, en revanche, manifestent surtout un pouvoir évocateur permettant de concentrer l'attention dans un champ visuel réduit sur une information qui s'éparpille dans le texte. Elles relèvent donc davantage de l'art que de la logique. Il n'existe pas de règles de bonne formation. Elles ne sont ni justes ni fausses, mais plus ou moins compactes et suggestives (voir Gentilhomme 1985).

Sans permettre de décider de façon péremptoire dans tous les cas, ce test autorise un premier déblayage du terrain.

Ainsi **N** (ensemble des nombres naturels), fonctionne dans $x \in \mathbf{N}$. À l'opposé, C.Q.F.Q., reste une abréviation. Tandis que l'on rencontre, dans des publications actuelles, PGCD et PPCM, sans points abréviatifs, utilisés dans des formules arithmétiques comme opérateurs :

$$PPCM(6,14) = 42, PGCD(6,14) = 2$$

Dans l'exposé de la théorie des capacités, créée par le mathématicien contemporain Gustave Choquet, on trouve des formules du genre $cap(K_1 \cup K_2) = cap\ K_2$, qui montrent que « cap » a dépassé l'étape abréviation pour devenir un symbole au sein de cette théorie.

Comme autre domaine en expansion de symboles issus d'abréviations, mentionnons les équations aux dimensions des physiciens (par exemple : $F = MLT^{-2}$) et surtout les nombreuses unités de mesure[18] qui rentrent dans des formules sans s'accorder en nombre et sans point d'abréviation :

$5cm \times 2\ cm = 10cm^2$, c (vitesse de la lumière) = 299792,458km/s.

Les notations telles que : 4 cm.17 ou 4,17 cms sont déclarées vicieuses.

Remarquons que l'origine linguistique peut être plus ou moins présente dans les symboles d'unités de mesure. Il est curieux de constater que, dans certains pays, la graphie du symbole subit une réécriture et qu'on peut parler de véritables calques, au sens où l'entendent les lexicologues. Ainsi, en russe, avant d'être abrégée, l'unité est traduite, puis notée avec des caractères cyrilliques. Ainsi : W (watt) est transcrit вт (ватт), rad/s (radian par seconde) – рад/сен (радиан в секунду), P (poize, viscosité dynamique) – п3 (пуаз), μV (microvolt) – мкв (микроволт), etc.[19] Cette particularité semble, d'ail-

18. Pour une documentation quasi complète, voir *Précis, Unités et grandeurs*, R. Quatremer et J.-P. Trotignon, Afnor, 3e éd. 1981.

19. Nos exemples sont tirés de l'ouvrage russe : Burdun, 1960, *Unités des grandeurs physiques*. Voir également : Gentilhomme, 1964, *Manuel de russe à l'usage des scientifiques*, pp. 622-623.

leurs, en voie de disparition, du moins dans les textes technoscientifiques, sous l'influence des publications internationales.

En chimie, les symboles d'éléments simples sont issus des noms (souvent latins) qui leur ont été attachés par la communauté scientifique : Az = N (azote, nitrogène), Mo (Molybdène), K (potassium, lat. Kalium)...

Dans un texte rédigé en style télégraphique (par exemple, un pense-bête, des notes d'étudiants), ces symboles peuvent « usurper » la place de mots à part entière, bien que l'information véhiculée soit plus riche, car les symboles représentent non seulement le corps simple mais, plus précisément, une certaine quantité, un atome-gramme de ce corps ; et, pour les formules (par exemple, $KMnO_4$, permanganate de potassium), une molécule-gramme du composé. (On peut entendre ou lire des énoncés comme : *les cristaux de $KMnO_4$ sont colorés en rouge pourpre*)

Bien d'autres motivations sont susceptibles d'intervenir pour la création de symboles. Le problème général dépasse notre présent propos.

Symboles générateurs de termes

Inversement, les symboles peuvent intervenir dans la morphologie des termes. Le cas où un symbole seul produit un nom propre à part entière n'est pas fréquent, du moins dans un style soutenu. Toutefois, mentionnons, par exemple, *le test du* (*la loi de*) *khi-deux* des statisticiens, la *loi gamma* (γ), voire le célèbre π, nombre transcendant, historiquement lié au cercle, qu'on ne nomme pratiquement que par la lettre grecque[20] ; on trouve même les *x(s)* pour dire *les inconnues.*

En revanche, on trouve de nombreux termes composés (figés ou semi-figés), pré- ou plus rarement postfixés avec des symboles : *radiation* α donne α-*irradié,* λ-*opérateur, N-illon* (10^{6n}), anneau $Z(\sqrt{a})$...

En chimie, de longues suites de radicaux et de symboles calquant, conformément à certaines conventions internationales, des formules développées, sont considérées comme des termes et figurent dans les textes ainsi que dans les dictionnaires spécialisés. Ces termes hybrides sont destinés plus à être « vus » qu'à être « entendus » ; aussi, la langue naturelle reprenant son droit, les remplace-t-on souvent par des synonymes plus compacts et linguistiquement maniables, mais dont la forme cesse d'être éclairante *a priori* pour la détermination de la structure du corps considéré[21].

Comparer : $(CH_3)_2CHCH_2COCH_3$ ou méthylisobutylcétone, encore dicible, et diméthyl 1', 1' butyl 7 diméthyl 1", 1" pentyl 7 tridécane

Aussi préfère-t-on se servir, pour l'usage courant, de termes plus compacts : *cholestérol*, ou d'abréviations : *A.D.N.*

Terme – unité de fonctionnement linguistique

Une terminologie, avons-nous dit, doit répondre à une double sollicitation : d'une part aux besoins de la discipline en cause, d'autre part aux impératifs inhérents à toute communication.

20. *Numéro spécial* π, Supplément au Petit Archimède, Mai 1980, 289 p.
21. Voir *op. cit.*, note 3, chap. *La correspondance voco-structurale,* p. 165 et sq.

Si l'on extrapole à la terminologie le principe du double fonctionnement – syntagmatique et paradigmatique – posé par Saussure, repris et développé par l'École structuraliste[22], un signe-terme, comme n'importe quel autre signe-mot, doit être considéré de deux points de vue complémentaires.

En effet, un terme immergé dans un discours réalisé dans diverses situations (exposé de haute tenue, vulgarisation, pédagogie, néologie heuristique, discussion familière entre pairs...) entretient, dans les discours mixtes, des relations sémantico-syntaxiques non seulement avec des signes-mots, dits usuels, et, entre autres, avec des « termes larvés », mais également avec des signes-symboles et suites de symboles (formules).

Termes larvés, « faux amis » terminologiques

Par « **termes larvés** », nous entendons des mots qui, à première vue, semblent appartenir au vocabulaire commun, mais qui, dans la discipline considérée, acquièrent un contenu particulier, non explicitement défini ; on ne le saisit qu'à la suite d'une pratique plus ou moins prolongée de la discipline. Citons, pour fixer les idées, des mots tels que : *exister, possible, nécessaire, vrai, absurde, probable, quelconque, remarquable, démontrer, montrer, prouver, établir, développer, expliquer, justifier, résoudre, considérer, définir, construire, comparer, tendre vers...* sources de malentendus et de dys-compréhensions chez des apprenants non aguerris. Ils jouent un rôle de « **faux amis** » difficiles à cerner et qui, bien que sémantiquement apparentés aux signes-termes de bon aloi[23], sont rarement répertoriés dans les lexiques terminologiques et ne figurent guère sur des fiches terminologiques.

Que signifie, pour un apprenant non aguerri : *triangle quelconque* ? Veut-on dire que le triangle n'est ni isocèle, ni rectangle ? Ou bien cela signifie qu'il ne faudra pas tenir compte dans les démonstrations de ses particularités, même s'il est isocèle ou rectangle ? Paradoxalement, un triangle isocèle ou rectangle peut néanmoins apparaître comme quelconque.

L'énumération des *points remarquables* d'un triangle s'arrête-t-elle à ceux qu'on a étudiés à l'école (*orthocentre, centre de gravité, centres des cercles circonscrits, inscrits et exinscrits*) ? Dans les ouvrages spécialisés, on en cite des dizaines : *points de Lemoine, de Brocart, de Gergonne, de Fermat, de Nagel, de Steiner, d'Euler, de Toricelli, de Napoléon...* tous plus remarquables les uns que les autres, et leur nombre n'est pas arrêté. Quelle information apporte-t-on en disant qu'un point est remarquable ? Quand les points cessent-ils d'être remarquables ?

D'une façon générale, comment interpréter l'antonymie *quelconque* contre *remarquable*, lorsque ces mots s'appliquent à des entités mathématiques comme : nombre, point, droite, plan, triangle, conique... ? Jacques Lobczanski pose le problème curieux (APMEP n°349, 1985, pp. 103 et sq.) : *Comment réussir le triangle quelconque ?* et poursuit le raisonnement (APMEP n°351, 1985, pp. 911 et sq.) pour aboutir à la conclusion paradoxale qu'il existe trois triangles les plus quelconques qui deviennent, de ce fait, des triangles remarquables. Ainsi *quelconque* et *remarquable* ne font que décrire l'attitude du locuteur vis-à-vis de l'entité considérée. Dès qu'on fixe son attention sur une entité réputée *quelconque*, elle cesse de l'être pour devenir *remarquable*. Ceci étant, *quelconque* acquiert une acception nouvelle définie par un calcul des probabilités.

Faux amis, déviances, abus sémiologiques

On peut se demander si le problème des faux amis se pose à propos des symboles.

22. Oswald Ducrot, 1968, *Le structuralisme en linguistique.*

23. Voir Gentilhomme, 1992a, p. 62. Robert Blanché, dans *L'axiomatique*, 1970, rappelle la distinction que faisait en son temps le géomètre Joseph Diaz Gergonne, dans *Essai sur la définition*, entre les définitions *explicites* et *implicites*. Le problème n'est pas nouveau, cependant la solution pédagogique reste à trouver, ce que souligne l'expression *terme larvé*.

Certes, il n'est pas impossible que, dans un texte scientifique ou technique, l'auteur utilise des symboles usuels avec un sens spécifique, différent, défini sans doute quelque part dans sa contribution et que le lecteur, qui n'en a pas eu connaissance ou qui l'a oublié, soit perturbé à cause de ses propres souvenirs.

Par ailleurs, la langue courante emprunte aux sciences, en guise d'abréviations, certains symboles en leur attribuant un contenu *ad hoc*. Tel est le cas des signes « = »,« + », « x », « < », « / » (barre de fraction), plus récemment : « $\exists$ », « $\subset$ », « $\cap$ »,« $\Rightarrow$ », etc. Sans doute, ces emplois sténographiques, plus ou moins farfelus ou raisonnables, ont quelque chose à voir avec les emplois codifiés dans les disciplines d'origine. Face à une censure rigoriste, ils n'en constituent pas moins une déviance ; à la limite, un « **abus sémiologique** ».

Que penser de certains locuteurs naïfs qui s'imaginent que, ce faisant, ils adoptent un « langage scientifique » ? L'expression figée, inscrite en langue : *démontrer par a + b*, c'est-à-dire, démontrer de façon rigoureuse, dénonce cette naïveté.

En ce sens on peut parler, métaphoriquement, de l'existence de **faux amis symboliques**.

Par ailleurs, hors discours en « langage technoscientifique », chaque signe-terme s'oppose, selon des modes divers, à d'autres termes simples ou complexes, considérés individuellement ou groupés en sous-ensembles et même, le cas échéant, à des signes-symboles, porteurs d'une information codifiée par l'usage (symboles « noms propres »).

Tel est le cas, en mathématiques et en physique, de π = 3,14159... (connu grâce aux ordinateurs avec plusieurs millions de décimales), e = 2,7182818... (base des logarithmes népériens), i. et j (nombre imaginaire, racine carrée de -1, noté de façon différente par les mathématiciens et les physiciens pour éviter la confusion avec l'intensité du courant électrique), $N = 6{,}0221 \times 10^{23}$ dit nombre d'Avogadro (nombre de molécules dans une molécule-gramme), h la constante de Planck, c la vitesse de la lumière, pH le potentiel-hydrogène...[24]

Un terme est susceptible de s'intégrer à diverses classifications, de nature linguistique, technologique ou hybride. On parlera de synonymie, d'antonymie, d'homonymie, d'isonymie de pantonymie, etc.[25]

Notre **hypothèse forte** sera que les cotoiements syntagmatiques et paradigmatiques des symboles avec des unités de langue (mots, termes) ne restent sans une certaine influence réciproque.

Avant de préciser et de justifier cette hypothèse, il convient de rappeler quelques propriétés des uns et des autres.

Sens et référent

En suivant la tradition (Aristote... Saussure... Frege[26]...), un **signifiant** (image mentale relevant de l'ouïe ou de la vue) est associé à un **signifié**, ou sens, renvoyant à un référent (objet, processus ou concept), sens que le lexicologue cherche à saisir et à rendre explicite, avec plus ou moins de précision, moyennant une **définition lexicologique**.

Il est important de souligner que le sens est supposé **exister indépendamment**

24. Voir liste officielle des symboles physiques : Quatremer et J.-P. Trotignon,1990.
25. Pour une analyse linguistique de ces notions, voir : John Lyons 1970, Bernard Pottier 1974, Robert Martin 1976, Igor Mel'čuk, 1984, 1988 et 1992. Pour des exemples concrets, consulter Bertaud du Chazaud, 1992.

de la volonté du lexicologue. C'est une donnée sociale. Signifiant et signifié sont liés, de façon **arbitraire**, bien que généralement **motivée**, comme les deux faces d'une pièce de monnaie.

Hypothèses de travail

Se posent aussitôt des questions d'ordre pragmatique : que faut-il entendre par signifié, pour les signe-mot, -terme et -symbole, comment l'identifier, le saisir, le cerner, le distinguer des autres signifiés et comment, en l'occurrence, interpréter le phénomène de polysémie ?

Plusieurs prises de position s'affrontent. Il ne nous appartient pas ici d'en discuter le bien-fondé.

Rappelons, cependant, que l'existence même d'un sens attaché à un mot hors contexte (à un terme) est nuancée par certains, carrément niée par d'autres.

D'abord, il est bien connu qu'en contexte-situation le contenu peut varier du tout au tout.

Ainsi le philosophe-linguiste P.P. Strawson[27] admet l'existence de trois sens pour un même énoncé. Considérons-en un adapté à notre propos : *branchez le rhéostat !*

1. Hors contexte-situation, *rhéostat* indique une information technique précise pour un électricien, le verbe *brancher* également ; d'où un premier sens, dit « littéral », exprimable dans d'autres langues avec des termes appropriés.

2. Cependant, on ignore qui a donné l'ordre, à qui, à quel moment, et de quel rhéostat il s'agit. En contexte-situation, ces questions trouvent leur réponse. Notamment, le rhéostat désigne un appareil bien déterminé, réalisé d'une certaine façon (alliage de la résistance, forme du support, etc.) et l'action se passe en un lieu connu, à un certain moment, etc. *Rhéostat* et *branchez* se chargent d'une information bien plus riche. D'où un sens dit « situationnel ».

3. En fait, le locuteur voulait simplement rappeler à un étudiant distrait : vous négligez les instructions du montage, vous risquez un court-circuit. Quel contenu attacher dans ces circonstances aux mots en question ? D'où un troisième sens dit « complet ».

Notre étude porte principalement sur le sens littéral.

Il n'est pas exclu que, dans certains de leurs emplois, on puisse associer aux symboles un sens plus ou moins riche en fonction du contexte-situation. Ainsi, un auteur qui n'utilise que des logarithmes de base 2 peut se contenter du symbole Lg, sans rappeler à chaque fois cette base, ce que ne peut faire un auteur qui fait appel à des logarithmes de base différentes. Cependant, *a priori*, on voit mal que ce symbole puisse signifier dans un exposé dogmatique : *vous vous êtes trompé de base, recommencez votre calcul* ! ou n'importe quel autre acte illocutoire.

D'autres auteurs, comme Hilary Putnam[28] mettent en doute la scientificité de la notion de sens comme manquant de stabilité. Retenons, notamment, l'idée du « sens par délégation ». De nombreux locuteurs sont capables d'utiliser correctement le terme *rhéostat* sans être en mesure, pour autant, de le définir, ne disposant que d'une vague représentation (objet servant à quelque chose en électricité, voire plus évasivement en science), avec une précision variable d'un locuteur à l'autre. Ce faisant, selon l'auteur, ils s'abritent, en quelque sorte, derrière l'autorité des spécialistes « qui, eux, savent exactement de quoi il retourne ». Ainsi, sens et locuteur finissent par former un couple indissociable : autant de locuteurs, autant de sens.

26. Voir Gottlob Frege, 1871, *Écrits logiques et philosophiques*, pp. 99 et sq.
27. Voir P. P. Strawson, 1964.
28. Voir Hilary Putnam 1990.

> Le clivage entre sens et référent s'estompe en praxématique[29], qui substitue au « signe saussurien », le praxème ou « couplage entre la forme du réel et une forme du langage ».

Nous nous plaçons dans la perspective terre-à-terre d'un lexicographe qui cherche à rédiger un article de dictionnaire contenant l'information dont l'utilisateur a besoin et à élaborer un système d'adressage permettant une consultation aisée.

Ne serait-ce que par nécessité rédactionnelle, la même forme étant susceptible d'être porteuse d'informations variées, celles-ci devront être dégagées, répertoriées, isolées, et classifiées. Il se trouve que des divergences apparaissent selon les auteurs, mais non sans une concordance approximative globale.

Nous admettons comme hypothèse de travail raisonnable (sans nous poser de questions fondamentales sur l'unicité du signifé et la multiplicité des effets de sens possibles) l'existence du phénomène de polysémie. Les divergences de présentation sont imputables à la nature même de la langue, relevant du problème général de la discrétisation d'un continum. Nous en prenons acte, tout en accordant notre préférence à certaines présentations en fonction de la finalité du dictionnaire.

Réalité contraignante et droit à l'« inventivité »

Quoi qu'il en soit, la langue se manifeste comme une réalité hypothétique qui se laisse appréhender à travers la parole, entre autres. Nous insistons particulièrement sur le fait que le lexicologue ne peut décider de son propre chef qu'un signifié (sens) donné doit s'associer à tel signifiant (forme-support) de son choix, plutôt qu'à un autre et vice-versa ; pas plus qu'un physicien-expérimentateur ne saurait décider selon son gré qu'un métal donné doit avoir telle masse volumique, telle résistivité électrique et qu'il doit fondre à telle température et à telle pression, ou encore, qu'un chimiste-expérimentateur (non prestidigitateur) ne peut imposer une couleur arbitraire au précipité d'hydroxyde de cuivre.

En ce sens, l'activité du lexicologue peut être qualifiée de « prospective ».

Comme nous le verrons, il n'en est pas tout à fait de même pour un signe-terme et encore moins pour un signe-symbole.

Sans doute, un terminologue (ou une commission terminologique) chargé de dresser un lexique s'en réfère à l'usage instauré dans la discipline en cause. Toutefois, dans une certaine mesure, sa tâche consiste également à normaliser et même à proposer des dénominations répondant mieux aux besoins des spécialistes concernés, mieux adaptées à la situation sociale, voire même politique (voir la lutte contre l'envahissement par des termes étrangers).

---

29. Voir Paul Siblot 1990.

Dans le Bulletin de l'APMEP n° 357, 1987, p. 52, la commission du second cycle explique certaines erreurs aux examens par une mauvaise compréhension du mot *comparer*.

Quelques termes larvés ont tout particulièrement attiré l'attention des pédagogues, des épistémologues et des historiens des sciences qui leur consacrent des recherches importantes. Voir, par exemple, les articles de l'APMEP Nouvelle rubrique de la commission « Mots » : *Mots flous*, 1992, n° 384, pp. 339-344.

Sur un plan plus général, voir : *Faits de langues. Motivation et iconicité*, recueil d'articles, présentation de L. Danon-Boileau, 1993.

L'activité néologistique du chercheur et du technicien (ou de leur communauté) fait davantage appel à l'imagination et à l'« **inventivité** ». Pour exprimer ce qu'ils ont à exprimer, ils sont obligés d'inventer des termes et des symboles nouveaux appropriés. Certes, la néologie n'est jamais complètement aléatoire, ni même libre. Elle reste soumise à des contraintes diverses de motivation et de systématisation que nous n'examinerons pas ici.

Concluons cependant qu'entre les tâches du lexicologue, du terminologue et du chercheur ou technicien, s'établit une progression allant de l'observation et de l'analyse d'un donné, jusqu'à l'élaboration inventive, quasi libre, d'un mode d'expression spécifique.

Comme nous le verrons, ce dernier caractère a un impact encore plus fort sur les symboles.

Synonymie linguistique et terminologique

> Il faut distinguer le sens de la dénotation. « 2^4 » et « 4.4 » ont bien la même dénotation mais pas le même sens. [...] Nous voudrions dire dans le cas présent que ces expressions ne contiennent pas les mêmes pensées.
>
> Frege, *op. cit.* p. 89.

Les termes : *spirale de Cornu, radioïde aux arcs* et *clothoïde* désignent la même courbe (issue des travaux sur la diffraction puis utilisée dans les dessins des bretelles d'autoroutes). Ces trois termes sont-ils des synonymes parfaits ?

La *quadratrice d'Abdank-Abakanowicz* a pour équation cartésienne

$$y = {}^1/_2R^2 \arcsin(x/R + {}^1/_2x)\sqrt{(R^2-x^2)}.$$

Peut-on considérer que le terme et l'expression la *courbe d'équation...* sont sémiologiquement synonymes[30] ?

En chimie, l'*acide prussique* (appellation obsolète, mais non totalement abandonnée, liée à son extraction du bleu de Prusse) et l'*acide cyanhydrique* (combinaison du cyanogène et de l'hydrogène, nommée ainsi par analogie avec les acides halogénhydriques FH, HCl BrH, IH) désignent la même substance de formule HCN. Cependant, les deux termes et la formule ne seront pas acceptés de la même façon dans tous les textes pour des raisons stylistiques.

L'idée de synonymie repose sur au moins deux critères, à savoir l'identité de contenu et la cosubstituabilité dans certains contextes.

En règle générale, nous admettrons que sens et référent constituent deux entités

30. Ces exemples sont tirés de la revue du Palais de la découverte, n° spécial 8, *Courbes mathématiques*, 1976. On y trouve de nombreux autres exemples intéressants pour notre propos.

distinctes du contenu. Nous distinguerons donc la **synonymie linguistique** (concernant le sens) de la **synonymie terminologique** (concernant le référent), en faisant abstraction, pour l'instant, des connotations sociales ou personnelles.

Ainsi, *cercle des neuf points* (cercle passant par les pieds des hauteurs, par celui des médianes et par trois autres points remarquables du triangle) et *cercle d'Euler* (cercle étudié particulièrement par le mathématicien suisse Euler) ont le même référent et sont de parfaits synonymes terminologiques. Ils ne le sont pas nécessairement du point de vue linguistique, car porteurs de sens différents : pour le nommer, on se réfère, d'une part, à un personnage historique, *Euler* ; d'autre part, à des éléments géométriques, *neuf points*[31].

Il y a coïncidence d'information, et même cosubstituabilité des deux signes-termes dans un contexte géométrique (moyennant des précautions définitionnelles), mais pas nécessairement dans un contexte historique. Les deux termes ne sont pas des synonymes linguistiques. La synonymie n'a pas la même portée lorsqu'il s'agit de signe-mot ou de signe-terme.

> Le signe polysémique « = », dans une de ses multiples significations, veut dire que les expressions situées à gauche et à droite sont deux désignations (différentes) d'une même entité. Au sens mathématique et logique, il a donc sa place légitime entre deux signes-termes :
>
> *cercle des neuf points* = *cercle d'Euler*
>
> Plus généralement, Francis Reynes remarque dans *Langage, synonymie, et démonstration* (*Bulletin de l'APMEP*, n°341,1981, p. 840): « En première approximation, on peut dire que la synonymie est aux phrases ce que l'égalité est aux désignations d'objets ».
>
> Il faudrait nuancer beaucoup pour s'autoriser à placer, en toute rigueur, le signe « = » entre deux signes-mots synonymes. Voici encore quelques exemples.
>
> Les noms des corps simples et, par suite, leurs symboles peuvent être l'enjeu de fierté nationale. (*Tungstène* rappelle l'origine suédoise *tungsten*, ou pierre pesante, tandis que *Wolfram* vient de l'allemand). D'où un effet connotatif sur certaines personnes, la plupart des usagers ne se posant pas la question.
>
> Le contexte peut également imposer des réserves. Comme nous l'avons vu, « i » et « j » dénotent tous les deux le nombre imaginaire, $\sqrt{-1}$; cependant « i » a sa place normale en contexte mathématique, « j » ayant une autre signification, mais non en contexte électrique pour raison d'ambiguïté (intensité d'un courant électrique).

Ainsi, en langue, deux synonymes (parfaits) – ayant rigoureusement le même sens et cosubstituables en tout contexte – sont rares. Les dictionnaires de synonymes font état, en fait, de parasynonymes (voir par exemple, Bertaud du Chazaud (1992)).

Au contraire, en terminologie, les synonymes cosubstituables et possédant le même référent sont assez courants, à condition de négliger les connotations sans importance pour la discipline.

> Il est clair que le nom propre *Abdank-Abakanowicz* rentrant dans la définition de la courbe, citée en exemple ci-dessus, produira un effet connotatif d'exotisme sur quelques lecteurs, effet auquel d'autres lecteurs resteront parfaitement insensibles.

31. Comme illustration, on cite souvent le même pont nommé différemment selon les riverains : *Pont de Kehl* ou *Pont de Strasbourg ;* ou encore : *le vainqueur d'Austerlitz, le petit caporal* et le *vaincu de Waterloo.* Cependant, un géomètre serait quelque peu surpris de lire, par exemple, *Le problème du vainqueur d'Austerlitz* pour *le problème de Napoléon-Mascheroni* (retrouver le centre d'un cercle donné en ne se servant que du compas).

Dans quelle mesure l'énoncé d'une définition est-il synonyme du terme défini, même si, logiquement, il peut se substituer à celui-ci ? Dans le modèle Sens-Texte[32] Igor A. Mel'čuk étend ce principe logique aux mots d'usage courant, en dépit de certaines rugosités de style qui en résultent.

Un parallélogramme, par exemple, peut être défini par une quelconque de ses propriétés caractéristiques. Ces définitions sont-elles toutes des synonymes du terme *parallélogramme* et synonymes entre elles ?

Les traités proposent plusieurs définitions très différentes les unes des autres pour présenter les coniques (ellipse, hyperbole, parabole).

Un ensemble infini est un ensemble dont le nombre d'éléments n'est pas limité ; ou un ensemble infini doit être équipotent à un de ses sous-ensembles.

Or, ces définitions ne peuvent s'insérer à n'importe quelle place du discours, car elles sont tributaires non seulement du développement logique de la discipline, mais encore des présupposés épistémologiques des auteurs. Ainsi, selon la place occupée dans le cursus déductif, un énoncé peut, selon le cas, acquérir le statut de définition ou de théorème. Il semble difficile d'accepter qu'un terme devienne synonyme (sans autres considérations) d'un énoncé à statut variable : définition ou théorème.

Pour en revenir aux symboles, une courbe peut être définie par ses équations en coordonnées cartésiennes orthonormées, par ses coordonnées polaires, voire bipolaires, par un couple d'équations paramétriques, par une équation différentielle, ou d'une façon dite intrinsèque (relation entre le rayon de courbure et l'abscisse curviligne) et j'en passe.

Ainsi pour la lemniscate de Bernouilli on a :

$$(x^2 + y^2)^2 = a^2(x^2-y),\ \wp = \sqrt{\cos 2\theta} \text{ et } MF.MF' = OF^2$$

Comment parler de synonymie dans ces conditions ?

Le problème soulevé mérite d'être approfondi, mais il dépasse de loin notre propos.

Concept, notion, fission du signifié[33]

Considérons un terme comme *radar* susceptible d'une définition rigoureuse, permettant d'opposer des appareils portant ce nom à d'autres appareils, des non-radars. Cependant si, au cours d'un entretien amical, un technicien du radar, s'adressant à son collègue, afin d'exprimer son admiration pour sa perspicacité, s'exclame : *tu as un radar au bout de ton nez !* il sera parfaitement compris, bien que le mot *radar* en ce con-

32. L'originalité du modèle transductif Sens-Texte, est, contrairement à d'autres modèles, de chercher à construire, à partir d'un texte donné, tous les textes susceptibles de véhiculer le même sens (traduction intralinguale). Voir : Mel'čuk 1984, 1988, 1992 et Gentilhomme 1992b.
33. Voir Gentilhomme 1982a et 1992.

texte ne puisse être remplacé par sa définition technologique. Nous en concluons que le signe-mot *radar* (contrairement au signe-terme technique *radar*) peut prendre (s'enrichir ou s'appauvrir) en langue, ne serait-ce que dans l'espace d'un instant, des contenus nouveaux, apparentés sans doute au concept technique, mais distincts.

Il se peut même que cette façon de s'exprimer obtienne un certain succès et s'inscrive en langue. Ceci étant, pour que le spécialiste-locuteur ait l'idée de ce néologisme sémantique, il faut bien que dans son esprit le contenu du mot *radar* ne soit pas réduit, *in æternum*, à sa définition technique stricte, qu'il y ait place – conjointement, mais en marge – pour un contenu souple, passible ultérieurement de définition lexicologique et que nous appelons **signifié notionnel** ($\mathbf{S_{notion}}$).

Inversement, des mots usuels, comme *union, intersection, application, élément, différence, réciproque, opposé, contenir...* utilisés dans un contexte mathématique deviennent des termes à part entière, associés à des définitions rigoureuses. Ceci étant, pour créer ces néologismes techniques, les spécialistes se sont inspirés du sens courant de ces mots.

C'est pourquoi nous posons l'hypothèse forte que l'observation du fonctionnement d'un signe-terme incite à dédoubler son signifié saussurien ($\mathbf{S_é}$).

D'une part, son contenu strict peut être celui d'une définition (explicite ou explicitable) que nous appelons le **signifié conceptuel** ($\mathbf{S_{concept}}$), qui seul, en principe, intervient dans le contexte technique, notamment au cours d'une démonstration. D'autre part, un terme peut évoquer toutes sortes de représentations, relevant de la métonymie et de la métaphore. En plus de son signifié conceptuel, le terme possède un quelque chose de plus : un **signifié notionnel** ($\mathbf{S_{notion}}$).

> La concomitance des deux signifiés se manifeste souvent dans la néologie sémantique où, à une forme déjà usitée, on attache un sens nouveau. En l'occurrence des mots d'usage courant, comme *courbure, torsion, translation...* deviennent des termes à part entière (voir Gentilhomme 1992, 1993). Souvent leur sens est décrit comme une spécialisation restrictive du sens commun. Pour peu que l'on connaisse ces termes à l'intérieur de la discipline concernée, on ne peut que nier une telle interprétation, comme il serait ridicule d'affirmer que *casserole* réflecteur électrique pouvant donner, au théâtre, des couleurs différentes, est une restriction du sens de *casserole* récipient pour la cuisine. Il s'agit en fait de polysémie allant parfois jusqu'à l'homonymie. Ce qui n'empêche nullement que l'un fasse penser à l'autre et réciproquement.

Nous résumons notre hypothèse par les formules-images,

pour le signifié : $\mathbf{S_é = (S_{concept}, S_{notion})}$

pour le signe-terme entier : $\mathbf{S_{terme} = (S_{forme} ; S_{concept}, S_{notion})}$

> À titre d'illustration, considérons quelques noms évocateurs de courbes : *Le scarabée, la besace, la conchoïde, le trifolium, la cardioïde, le bicorne, le limaçon de Pascal, la courbe du forçat, le trident de Newton ...* (pris dans *op. cit. Courbes mathématiques*). Il semble difficile de nier que dans l'esprit (voire dans l'inconscient) de celui qui les nomme, qui en étudie les propriétés, ne subsistent pas des traces du sens des mots courants utilisés pour les nommer. Certes, le signifié notionnel n'intervient pas dans le raisonnement. Un discours scientifique ne se réduit pas au seul raisonnement logique. La néologie y joue également son rôle. Pourquoi baptiser ces courbes de noms évocateurs ?

De nombreux autres exemples peuvent être cités dans d'autres disciplines.
Nous ne reprendrons pas ici les arguments sur ce thème développés par ailleurs (voir Gentilhomme 1992a et 1993).

Symboles, termes, mots

Jetons un coup d'œil rapide sur quelques propriétés des symboles (ou aux arrangements divers de symboles) relevant plus particulièrement de la sémantique, qui les opposent aux termes et aux mots ou, au contraire, les en rappprochent.

Nous n'aborderons pas les propriétés relevant davantage de la syntaxe.

Ainsi, on distingue souvent les symboles-constantes (π, e, N...) des symboles prédicatifs (log, tan...). Nous ne discuterons pas cette vaste problématique ici.

En revanche, dans certains de leurs fonctionnements, peut s'appliquer le phénomène que nous avons appelé **fission du signifié**. Autrement dit, outre leur signifié conceptuel, il est possible de leur associer un contenu notionnel ou connotatif.

Sans parler de certains préjugés comme le « 13 » portant malheur ou chance selon les pays (en Extrême-Orient, c'est le « 4 » qui est réputé néfaste, d'où l'insuccès de la 404 Peugeot) ; sans évoquer la discipline ésotérique appelée Numérologie, où il n'y a pas de nombre quelconque, chaque nombre étant chargé de sens magique, dans notre civilisation, certains symboles techno-scientifiques peuvent évoquer dans l'esprit du profane, les « mathématiques supérieures » (*sic.*), ou la soi-disant absconsité des « maths modernes ». Plus modestement, les « X » font penser à l'algèbre « subie à l'école », avec ses inconnues et ses équations.

Les symboles « C », « A », « P » rappellent aux professionnels des mathématiques outre leur double sens technique défini rigoureusement en combinatoire, à savoir : certaines « applications » appelées en l'occurrence : *combinaisons, arrangements, permutations,* ainsi que les assemblages mathématiques obtenus par ces applications, l'acte commun de combiner certains objets concrets, selon certaines conventions plus ou moins clairement exprimées, comme cela se pratique dans la vie courante.
Le symbole « ds^2 » évoque dans l'esprit d'un spécialiste plus que sa définition algébrique stricte en géométrie analytique ; logiquement, il sera associé au problème général de rectification d'une courbe et, par suite, aux diverses expressions du « ds^2 » selon le choix du système de représentation de la courbe, au calcul requis par l'intégrale en cause, etc.

« Homonymie symbolique »

Du point de vue **stabilité sémantique**, les symboles peuvent être opposés, grosso modo, selon deux tendances. Plus précisément, à l'intérieur d'un domaine circonscrit, certains symboles sont porteurs d'un contenu relativement stabilisé par l'usage, jouissent d'un large consensus de la part de la communauté des spécialistes et jouent, en quelque sorte, un rôle de « noms propres » pour les objets de pensée (constantes, opérations, relations) :

π, e (déjà cités), [], { }, $\emptyset$, <, +, /, $\Rightarrow$, $\cap$, $\int$ (en math.).

Néanmoins, ils sont susceptibles de changer de contenu dans un autre domaine : e (charge électrique d'un électron), N nombre d'Avogadro, F constante de Faraday en physique, mais atome gramme d'azote et de fluor en chimie, etc.

Ce phénomène relève de l'**homonymie symbolique**. On ne peut pas soutenir qu'il y a eu un dérapage sémantique suite à quelque figure de style. N atome-gramme d'azote et N nombre d'Avogadro n'ont rien de commun, ni en diachronie (on ne peut proposer un étymon commun comme, par exemple, pour *voler,* s'accaparer d'un objet, ou se déplacer dans l'air), ni en synchronie (le nombre d'Avogadro N ne résulte pas d'un dérapage sémantique de l'azote N, ni vice-versa).

La même lettre majuscule a été utilisée à deux fins distinctes, jadis, par certains auteurs disposant d'une certaine autorité et s'est maintenue à l'usage.

« Multisémie symbolique », disponibilité

En revanche, d'autres symboles, notamment les symboles littéraux, semblent **disponibles** pour désigner tout ce que l'on veut, mais de façon bien déterminée à l'intérieur d'un contexte donné, pour une durée connue. Préalablement, leur contenu doit être rendu parfaitement explicite. Les variations métonymiques, si faibles soient-elles, sont exclues. Toute inobservance de cette règle, qu'on pourrait baptiser : **« fidélité à la parole donnée »**, constitue une faute grave.

> On pourrait rapprocher ce phénomène de l'emploi que l'on fait de certains mots comme : *tipoter, schtroumpfer, bidule, machin, chose, truc,* etc., sauf que ces mots, en nombre très réduit, ne sont pas astreints à une discipline rigoureuse comme ci-dessus.

Sans doute, cette liberté de contenu est-elle tempérée par des habitudes, par des commodités mnémotechniques, par une certaine cohérence, mais l'auteur n'est nullement obligé d'obtempérer.

Souvent, sans que cela soit impératif, on désigne une entité par son initiale latine ou par son équivalent grec : p, P, π, (ou sa variante graphique ϖ), Π sont évocatrices pour désigner un plan, une probabilité, une pression, un paraboloïde... et apparaissent presque comme des abréviations. On appellera volontiers un cercle c, C, ou o, O, ω, Ω ; une rotation Rot, puis simplement R. Rien, cependant, n'interdit de changer de notations, à condition, bien sûr, de prévenir l'interlocuteur.

Par une sorte de **contamination alphabétique**, les lettres voisines désigneront un autre exemplaire de même catégorie : Q, R, S pour un plan, q = p-1 pour la probabilité complémentaire, γ, Γ au lieu de c, C pour un cercle.

Il est de tradition que x, y, z, jouent le rôle d'inconnues ou de variables, tandis que les premières lettres de l'alphabet servent plutôt de coefficients. Cependant, le cas échéant, il n'est pas interdit d'inverser les rôles.

Fétichisme de la forme.

Certains pédagogues estiment même utile de faire varier les notations pour lutter contre ce qu'on peut appeler « le fétichisme de la forme », autrement dit pour faire por-

ter davantage l'attention des apprenants sur le fond (notamment sur les relations), plutôt que sur la forme, cela, en bousculant d'une manière un peu brutale leurs habitudes, tout en s'appuyant sur la célèbre boutade de David Hilbert qui s'amusait à remplacer, par des noms d'objets courants (*bocs de bière, tables, chaises*) les dénominations traditionnelles des entités mathématiques comme *point, droite, plan.*

Soulignons le fait qu'il serait bizarre de se comporter de la même façon en langue, sinon par jeu ou par encodage cryptique, et de convenir, par exemple, qu'à la place de *carotte* on dira *chou-fleur,* à la place de *maison, président de la République,* etc.

Insistons bien sur la différence fondamentale entre **polysémie** et **multisémie**. Dans un dictionnaire, il est possible d'énumérer de façon raisonnable, disons pratique, les différents contenus que l'on peut attribuer *a priori* au lexème entrée de la rubrique. Ces contenus sont passibles de certaines variations, de certains effets de sens en contexte, que l'on devine en partie grâce aux définitions lexicographiques.

Rien de tel pour nombre de symboles. *A priori*, on ne peut leur en attribuer aucun. En revanche, en contexte, ils peuvent acquérir n'importe lequel. Ce ne sont que des réceptacles de contenu, vides, disponibles, bien qu'ayant, en un autre moment, été remplis par des contenus déterminés et ayant un souvenir plus ou moins marqué des contenus antérieurs.

Disons, par plaisanterie, qu'un verre ayant contenu du pétrole doit être bien nettoyé pour servir de coupe à champagne.

Ceci étant, tout effet de sens est rigoureusement interdit.

Arbitrarité faible et forte

Selon la tradition issue de l'enseignement de Saussure, le lien entre le signifiant et le signifié est, en un certain sens, arbitraire, bien que motivé par des raisons diverses (étymologiques, onomatopéiques, sociales, etc.)

Sans doute en est-il de même pour les symboles et leurs assemblages. Toutefois, la nature de l'arbitrarité ne se présente pas de la même façon.

Ainsi, plus particulièrement pour les symboles littéraux disponibles, l'auteur reste totalement libre, avec la contrainte expresse due à la parole donnée, avons-nous dit, de doter un symbole du contenu de son choix, sans avoir même à le justifier et sans qu'un collègue ait le droit de lui en faire le reproche au niveau théorique, même s'il n'approuve pas ce choix en son cœur.

« Hystérèse sémantique »

Tel ne semble pas être le cas dans les sciences humaines où les termes empruntés à la langue courante ou à d'autres théories restent plus ou moins chargés des significations antérieures, phénomène que nous appelons hystérèse sémantique, par référence au phénomène d'hystérèse connu dans la théorie de l'aimantation (l'aimant se souvenant, en quelque sorte, de ses états antérieurs).

Sans doute l'hystérèse sémantique est toujours présente en néologie, même dans les sciences dites dures, mais elle n'y a qu'une valeur « pédagogique », ce n'est qu'un facteur de facilitation de la communication. L'auteur ne transgresse pas un tabou déon-

tologique en oubliant, en gommant le passé sémantique d'une forme-support, sauf à en avertir l'interlocuteur (voire le rappeler ostensiblement) et à ne pas doter une même forme de deux contenus différents non identifiables par le contexte.

Pour mieux nous faire comprendre, risquons une boutade, à la façon du mathématicien Henri Lebesgue : Pourquoi baptisez-vous du nom connu « x » telle entité ? Vous risquez d'introduire la confusion dans l'esprit de votre interlocuteur. – C'est vrai, alors appelons-la *trouloulou* ou *tralala*, comme il vous plaira, en espérant que ces nouvelles appellations originales, *ad hoc*, n'induiront pas le trouble dans votre esprit (voir Félix 1974).

On peut donc envisager deux pôles extrêmes de l'arbitraire, que nous appelons **arbitrarité faible** plus normale pour les mots d'usage courant, et **arbitrarité forte** dont jouissent particulièrement les symboles disponibles. La situation des termes et des symboles-noms propres rejoint plus ou moins celle des symboles disponibles, la cohérence du système terminologique imposant ses propres contraintes.

« Anonymie symbolique », disponibilité totale

Il existe des symboles-noms propres, dits mutificateurs, qui rendent **anonymes** les variables qu'ils recouvrent. Expliquons-nous.

Soit la formule définissant le cosinus intégral :

$$\mathrm{Ci}(x) = -\int x \cos t \,/t.dt.$$

On aurait pu remplacer la lettre t par n'importe quelle autre, sans définition préalable : la lettre t ne figurant pas dans le résultat final, « elle n'a pas droit à la parole », elle reste « muette ».

Ainsi, certains symboles comme le signe intégral « $\int$ » sont capables, dans certaines conditions, de rendre « muets », dans le jargon des mathématiciens (nous disons « anonymes » dans le nôtre), d'autres symboles. Ces derniers perdent, en quelque sorte, leur identité. Qu'importe qui ils sont, quel est leur nom, tout ce qu'on leur demande c'est d'exister et d'effectuer une tâche conforme à la consigne donnée. Dans une métaphore hardie, on pourrait parler d'esclaves au service d'un maître.

Il ne semble guère qu'il puisse en être ainsi pour des termes ou des mots usuels. Pour simuler ce phénomène en langue, il faudrait imaginer un texte tel qu'on puisse remplacer un mot ou un terme donné par n'importe quel autre appartenant au même paradigme, sans que le contenu du message en soit modifié. Il n'est pas totalement exclu qu'une telle prouesse puisse être réalisée par un linguiste-acrobate, mais elle ne relève nullement du quotidien de l'hexagonal moyen.

Quoi qu'il en soit, ce phénomène semble lié à l'**arbitrarité forte** des symboles. La variable sémiologique ne représente rien en elle-même, elle n'est que ce que l'on veut qu'elle soit, elle est le jouet de l'arbitraire du maître ; n'importe quelle autre pourrait la remplacer, son nom n'a aucune importance, elle est anonyme, ce qui n'est pas tout à fait le cas d'un mot usuel, lourd de traditions, ou d'un terme, doté d'une définition stricte.

Il nous appartient, à présent, de voir par quel acte de parole s'acquiert le contenu d'un terme ou d'un symbole.

Performativité et acte de parole définitoire

> NOTATIONS. Pour un polytope, introduisons les notations suivantes :
> Ξ = ensemble des arêtes de P, $\alpha = \# \Xi$,
> $\sum_i$ = ensemble des sommets de P appartenant à i faces, $\sigma_i = \# \sum_i$,
> $\sum$ = ensemble de tous les sommets de P, $\sigma = \# \sum$,
> Φ_i = ensemble des faces de P qui ont i côtés, $\varphi_i = \# \Phi_i$
> Φ = ensemble de toutes les faces de P, $\varphi = \# \Phi$.
>
> Extrait de M. Berger, *op. cit.*, p. 122.

Le philosophe-linguiste anglais J. L. Austin appelle verbes performatifs les verbes dont l'énonciation revient à réaliser l'action qu'ils expriment et qui décrivent une certaine action du sujet parlant. Tel est le cas de *je promets, j'ouvre la séance,* parce qu'en prononçant ces phrases on fait justement l'action de promettre, d'ouvrir la séance. (c'est parce que je dis : « je jure telle chose », que « telle chose » est promise et qu'on peut m'accuser de parjure si cela s'avère faux).

En ce sens, dans les textes technoscientifiques, le mot *définition* possède un pouvoir performatif car, en le prononçant, on réalise l'acte de définir le terme concerné.

L'acte de parole définitoire peut être obtenu moyennant des marqueurs divers, en général, des expressions langagières :

Appelons, nommons, est appelé, s'appelle, introduisons les notations, soit, on dit que, désignons par, ou...

> Par exemple : ***Appelons*** *symédiane la droite issue du sommet du triangle...*
> ***Soit*** Δ *la droite parallèle à D et passant par A ...*
> *Le point de concours des trois axes radicaux* ***ou*** *(**appelé, dit**) centre radical...*
> ***Désignons par*** *C(n) la proposition :* $\forall\ \kappa (0 \leq \kappa \leq n \Rightarrow | u_\kappa | \leq 1 / n+1)$

La ponctuation peut aussi avoir un effet performatif, notamment les « : » et les parenthèses.

La valeur illocutoire des mot fixant leur propre statut au sein de la théorie comme : *hypothèse, présupposé, principe, conjecture...* s'apparente à celui de *définition.*

En règle générale, en prime occurrence (non en reprise ou en rappel), une assertion acquiert le statut d'hypothèse – plutôt que d'un présupposé, d'un principe ou d'une conjecture – à partir du moment où on la déclare être une hypothèse. On suppose, bien entendu, que cette décision reste compatible avec la théorie exposée, sinon la déclaration du statut relève de l'erreur.

Il n'en est pas tout à fait de même pour des marqueurs du genre : *théorème, conséquence, corollaire, lemme, exemple.* Si l'on a affaire à un théorème, à une conséquence, à un exemple... ce n'est pas parce qu'on a prononcé l'un de ces mots, mais

davantage parce que le statut de l'énoncé résulte du développement logique de la théorie.

Les symboles à caractère performatif sont peu nombreux.

Les logiciens utilisent un symbole performatif spécifique « $=_{déf}$ », prononcé [egalpardefinisjɔ̃].

Le symbole « = » contribue parfois à l'acte définitoire ; voir texte en exergue définissant les symboles situés à gauche de ce signe.

Il en est peut-être de même pour les quantificateurs « $\forall$ », « $\exists$ », ainsi que pour les symboles d'appartenance « $\in$ », « $\notin$ », dans certains contextes. La question reste à débattre.

Ouverture d'une problématique

Comme le lecteur l'a certainement remarqué, le problème de convergences, divergences, interférences entre les symboles, les termes et les mots de la langue commune est immense ; il ne peut être traité dans son ensemble par un chercheur isolé dont les compétences sont forcément limitées. Aussi, pour les quelques questions soulevées, les réponses n'ont été qu'esquissées. Néanmoins, nous espérons que cette « première approximation » de la problématique, donne envie à d'autres chercheurs de creuser dans cette voie, en dépassant les lieux communs et les préjugés qui abondent dans ce domaine.

Pour clore cette contribution, nous aimerions ouvrir la voie à une vaste problématique, à savoir l'application de la méthodologie des fonctions lexicales du modèle Sens-Textes (voir Mel'čuk, *op. cit.* 1984 ; 1988 ; 1992 et Gentilhomme 1992b), conçu en principe pour le vocabulaire courant, aux termes et aux symboles, comme ils apparaissent dans les discours technoscientifiques.

Les renseignements recueillis par toute une équipe de chercheurs sont consignés dans une base de données, le *D.E.C.* ou *Dictionnaire explicatif et combinatoire du français contemporain.*

D'ores et déjà, les quelques tentatives en matière de terminologies particulières, poursuivies dans ce sens (notamment sous forme direction de mémoires de maîtrises et de thèses, à la Faculté des lettres et sciences humaines de Franche-Comté) nous ont obligé à revoir et à repenser nombre de questions.

Fonctions lexicales Magn, Degrad, Func, Oper, Syn

Rappelons, à l'aide d'exemples tirés des discours technoscientifiques, ce qu'on entend par fonction lexicale (FL).

La qualification générale d'intensité s'exprime en langue de façon diverse, selon

le mot sur lequel elle porte, par des unités de langue plus ou moins standardisées, par des expressions figées ou semi-figées.

Ainsi pour parler d'une pluie de forte intensité, on pourra dire certes, *une forte pluie,* mais aussi *une pluie battante, une pluie diluvienne* (dans une situation de communication non technique)...

Pour informer qu'on a affaire à une forte différence de potentiel électrique, on dira *une tension élevée,* mais non **une tension battante, *une tension diluvienne* (sauf à vouloir transgresser délibérément l'usage normal).

Il se trouve que dans les textes spécialisés, certaines expressions sont privilégiées, d'autres sont exclues. Il est intéressant, par exemple pour un traducteur, de savoir comment les spécialistes s'expriment entre eux et de pouvoir trouver la réponse dans des dictionnaires appropriés.

Pour la *croissance* d'une fonction dans un contexte algébrique, on parlera de *forte croissance,* mais pas (ou moins fréquemment), semble-t-il, d'une *croissance intense, puissante, grande, importante.* Non que ce soit grammaticalement incorrect, mais parce que telle n'est pas l'habitude dans cette communauté scientifique. En l'occurrence, le premier adjectif qui vient sur le bout de la langue c'est *fort.*

Le lexicographe n'a pas à en discuter le bien-fondé, mais à en enregistrer l'usage dans une certaine communauté.

En empruntant aux mathématiques la notation fonctionnelle, on introduit la « fonction » **Magn**, portant sur l'argument lexical *croissance* et prenant la valeur lexicale *fort.* On écrit donc :

Magn(*croissance*) = *forte*

Il arrive que cette FL puisse s'appliquer (à des nuances près) à des symboles :

Magn(<) = «

Comme autre exemple, citons la fonction **Degrad** qui indique le verbe utilisé communément pour signifier qu'une entité perd ses qualités pertinentes d'un certain point de vue généralement admis. On écrit donc :

Degrad(*miroir*) = *se ternir*, **Degrad**(*roulement*) = *se gripper,*
Degrad(*conique*) = *dégénérer...*

Comment les spécialistes s'expriment-ils pour dire qu'un acide perd ses vertus d'acidité, qu'une balance cesse de donner des renseignement fiables, qu'un lubrifiant ne lubrifie plus correctement, etc. ?

À l'inverse, la FL **Prepar** envisage l'action d'apprêter un objet pour qu'il puisse rendre le service qu'on attend de lui :

Prepar(*lame de rabot*) = *affûter, aiguiser*

Dans un autre ordre d'idées, quel verbe utilise-t-on normalement pour parler, par exemple, d'oxydation. Que fait « linguistiquement » un acide vis-à-vis du corps sur lequel il exerce une action caractéristique ? Il s'ensuit d'autres fonctions lexicales. On pourra lire ou entendre :

L'oxydation s'est effectuée à pression constante, d'où :
Func(*oxydation*) = *s'effectuer, avoir lieu...*

Un acide attaque un métal, (métaphore qui n'est pas valable dans toutes les langues), d'où :

Oper(*acide*) = *attaquer, agir sur...*

Dans le modèle Sens-Texte, l'auteur envisage une soixantaine de FL simples, composables entre elles.

Notons que suivant que les spécialistes parlent ou non entre pairs, la façon de s'exprimer peut changer ; en outre, cela peut être un signe de reconnaissance entre gens du métier (il peut en résulter, notamment, une incidence sur le prix). À côté du vocabulaire technique et d'une certaine façon de l'utiliser, il peut y avoir un vocabulaire et une façon de parler pour non-spécialistes. Ce problème, dans le domaine du sport a été étudié par Galisson, 1978).

Il ne peut guère être question ici d'examiner toutes les FL, ni de voir comment elles sont susceptibles de s'appliquer aux termes et aux symboles dans des textes à caractère technique.

Nous avons déjà dit quelques mots sur la synonymie qui s'exprime sous forme de quatre fonctions : **Syn**, **Syn**$\supset$, **Syn**$\subset$, **Syn**$\cap$, selon qu'on envisage une synonymie exacte, plus large (hyperonymie), plus étroite (hyponymie) ou approximative (parasynonymie).

Les terminologues nous invitent à concevoir des FL formalisant les situations d'isonymie, de pantonymie et d'idionymie (Gouadec *op. cit.*). Jetons un coup d'œil sur la FL antonymie et sur ses variantes.

Antonymie, Anti, Contr, Conv

« Les antonymes sont des unités dont les sens sont contraires ; cette notion de *contraire* se définit en général par rapport à des termes voisins, ceux de complémentaires (mâle contre femelle) et de réciproque (vendre contre acheter). »

On trouvera une analyse finement élaborée dans Martin (1976).

Le D.E.C. introduit la FL : **Anti**, avec les nuances : **Anti**$\subset$, **Anti**$\supset$ et **Anti**$\cap$, comme pour la synonymie. La fonction **Anti** se combine souvent avec d'autres :

Magn(*température*) = *élevée, haute* vs **AntiMagn**(*temp.*) = *basse.*

Dans une analyse moins superficielle, l'antonymie pose de nombreux problèmes et recouvre parfois des situations très différentes.

Ainsi, *précéder* apparaît comme l'antonyme de *suivre,* mais aussi comme converse, c'est-à-dire permettant de décrire la même situation en inversant les participants sémantiques en cause :

x précède y contient la même information que *y suit x,*
Conv *(suivre)* = *précéder* <=> **Conv** *(précéder)* = *suivre*

La différence ne relève que de la présentation des faits.

La FL « terme contrastif » oppose deux lexèmes sans qu'on puisse évoquer à proprement parler l'antonymie.

Ainsi, dans un triangle on a : **Contr**(*sommet*) = *base.*
En revanche, s'il s'agit d'une montagne : **Contr**(*sommet*) = *pied.*
En parlant d'un ordinateur : **Contr**(*logiciel*) = *matériel.*

On peut même envisager pour l'antonymie et pour les FL apparentées des oppositions triangulaires. Dans la description d'un polyèdre, les termes *sommet, face* et *arête* s'opposent deux à deux, sans être des contraires.

Peut-on étendre ces considérations aux symboles ?

Symboles disponibles

Pour les symboles totalement disponibles, pris hors contexte, la question perd beaucoup de son intérêt, leur contenu n'étant pas fixé, mais défini par l'auteur. Des symboles comme « a » et « A » peuvent, selon le cas, se trouver en situation de synonymie, hyponymie, hyperonymie, antonymie ou n'entretenir aucune relation de cette nature.

Toutefois, comme nous l'avons signalé, s'instaurent des habitudes et persistent des traces du passé (phénomène d'hystérèse) ; donc, certaines lettres ont des affectations préférentielles qui les opposent à d'autres, phénomène que l'on peut classer sous la FL $\mathbf{Contr}_{\text{hystérèse}}$ dûment indicée, portant sur des ensembles :

$$\mathbf{Contr}_{\text{hyst}}\ (\{x, y, z\}) = \{a, b, c\}$$

Les premiers faisant préférentiellement office de variables ou d'inconnues (rôle corroboré par des expressions semi-figées du genre *porter plainte contre X*) et les secondes de constantes. Des formules, telles que « $y = ax^2 + bx + c$ » ou « $ax + by + c = O$ », gravées dans les mémoires (voire dans l'inconscient) des écoliers qui les ont subies contribuent à maintenir la tradition.

À un niveau plus avancé, apparaît l'opposition conceptuelle, marquée par l'opposition littérale, entre nombres réels (x, y) et complexes (z = x + iy), d'où : $\mathbf{Contr}_{\text{hyst.}}$ $(\{x, y\}) = \{z\}$.

En puisant dans les manuels scolaires, il est facile de trouver de nombreux autres exemples.

Symboles-noms propres

Pour ce qui est des symboles-supports d'information relativement stable, comme les symboles-noms propres, les FL peuvent également être mises en œuvre.

Ainsi pour « ⟨⟨ » (*très inférieur à*) et « ⟩⟩ » (*très supérieur à*),
on peut écrire : **Conv**(⋙) = ⋘ et **Conv**(⋘) = ⋙.

Plus généralement, la FL **Conv** s'apparente à la notion de fonction réciproque, moyennant certaines précautions.

$$\textbf{Conv}(\tan x) = \arctan x \text{ (avec la restriction } -\pi/2 < x < +\pi/2)$$

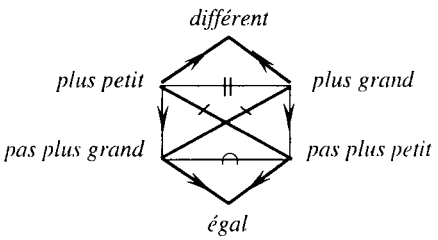
Mais quelle analogie établir entre les faits de langue et de sémiologie :

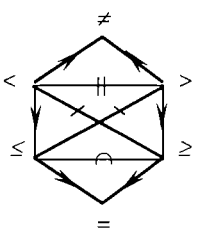
$$\tan^{-1} x \text{ fonction inverse et } -\tan x \text{ fonction opposée ?}$$

Le problème de l'antonymie et de ses variantes mérite d'être précisé et développé en sémiologie des discours technoscientifiques hétérogènes.

Hexagone de Blanché

L'hexagone de Blanché (1969) (développement du carré d'Apulée) s'applique aussi bien à certains « paradigmes » terminologiques qu'à certains « paradigmes symboliques » :





Une étude plus approfondie nous entraînerait trop loin ; nous nous contentons donc, encore une fois, de mentionner le problème.

Conclusion

L'objectif de cette étude était de comparer, en se plaçant du point de vue d'un lexicologue, les termes et les symboles, en se fondant sur leur comportement dans des discours technoscientifiques hétérogènes. Le sujet étant trop vaste et inabordable dans son entièreté par un seul chercheur, nous nous sommes limité à un domaine technoscientifique restreint.

Même dans ces conditions, au fur et à mesure de la progression de notre recherche, il s'est avéré qu'une analyse en profondeur dépassait le cadre d'un modeste exposé, de sorte que contraint et forcé, nous nous sommes contentés de n'effleurer que quelques aspects de la problématique, de ne poser que quelques questions et de n'avancer prudemment que quelques hypothèses.

Grosso modo, il semble qu'un terme occupe une position intermédiaire entre le

mot d'usage commun et divers types de symboles ; qu'il existe, en outre, une certaine influence réciproque entre termes et symboles, tant au niveau formel (morphologiques : abréviations, calques, termes hétérogènes), qu'à celui du contenu.

Le contenu sémantique d'un terme ou d'un symbole est double : d'une part, celui d'une définition rigoureuse, rigide, le seul valorisé dans certains emplois, donnant lieu à un **signifié conceptuel** ; d'autre part un sens plus ou moins vague, permettant, notamment, des figures de style, passible seulement d'une définition lexicographique, ou **signifé notionnel**, jouant un rôle important dans le processus complexe de l'heuristique.

Cette **fission du signifié** a de multiples conséquences, par exemple, l'opposition entre la **synonymie terminologique** et **linguistique**. L'homonymie a une portée spécifique. La polysémie s'oppose à la **multisémie**. La **disponibilité sémantique** des symboles poussée à l'extrême incite à s'intéresser à la propriété d'**anonymie**, etc.

Par ailleurs, nous avons relevé la différence de nature entre l'**arbitrarité forte** du symbole ou du terme par opposition à l'**arbitrarité faible** du vocabulaire courant, ainsi que du rôle de la motivation, en particulier à celui de l'**hystérèse**.

L'acte de parole définitoire acquiert une importance fondamentale (**performativité néologique**).

Pour finir, nous avons cherché à ouvrir tout un champ de recherche prometteur – l'application des **fonctions lexicales** du modèle transductif Sens-Texte à l'étude des termes et des symboles, qui, d'ores et déjà, nous a apporté des résultats originaux, mais que, dans cette contribution réduite, nous ne pouvons qu'esquisser.

Références

APMEP *Bulletin de l'Association des Professeurs de Mathématiques de l'Enseignement Public, de la Maternelle à l'Université,* Paris, 26, r. Duméril, 75013.

AUSTIN, John Langshaw (1970) : *Quand dire, c'est faire,* (éd. origin. *How to do Things with Words,* 1962), Paris, Seuil, 187 p.

BERGER, Marcel (1990) : *Géométrie,* Paris, Nathan, vol. 2, 541 p.

BERTAUD DU CHAZAUD, Henri (1992) : *Dictionnaire de synonymes et contraires,* Paris, Le Robert, « les usuels », 768 p.

BLANCHE, Robert (1969) : *Structures intellectuelles,* Paris, Lib. philosophique Vrin, 149 p.

BLANCHE, Robert (1970) : *L'axiomatique,* Paris, PUF, 110 p.

BURDUN, G. D. et Г.Д. Бурдун (1960) : Единицы физических величи, Moscou, СТАНДАРТГИЗ, 115 p.

CHEVALLARD, Yves (1985) : *La transposition didactique, du savoir savant au savoir enseigné,* Grenoble, la Pensée sauvage, 126 p.

Courbes mathématiques (1976) : Revue du Palais de la découverte, numéro spécial 8, juillet, 167 p.

DAGOGNET, François (1969) : *Tableaux et langages de la chimie*, Paris, Le Seuil, 221 p.

DANON-BOILEAU, Laurent (présentation de) (1993) : *Faits de langues. Motivation et iconocité*, recueil d'articles, 284 p.

DUCROT, Oswald (1968) : « Le structuralisme en linguistique », *Qu'est-ce que le structuralisme ?*, Paris, Le Seuil, pp. 13-96.

ECO, Umberto (1973) : *Le signe. Histoire et analyse d'un concept*, Bruxelles, Labor, trad. franç. 1988, 283 p.

FELIX, Lucienne (1974) : *Message d'un mathématicien. Henri Lebesgue, pour le centenaire de sa naissance*, Préface de S. Mandelbrojt, Paris, Blanchard, 259 p.

FREGE, Gottlob (1971) : *Écrits logiques et philosophiques*, (trad. et introd. de Claude Imbert), Paris, Seuil, 234 p.

GALISSON, Robert (1970) : *L'apprentissage systématique du vocabulaire, livres du maître*, Paris, Hachette-Larousse, 127 p.

GALISSON, Robert (1978) : *Recherches de lexicologie descriptive. La banalisation lexicale. Contribution aux recherches sur les langues techniques*, Paris, Nathan, 432 p.

GALISSON, Robert (1983) : *Des mots pour communiquer. Éléments de lexicométhodologie*, Paris, CLE international.

GENTILHOMME, Yves (1964) : *Manuel de russe à l'usage des scientifiques*, Paris, Dunod, 659 p.

GENTILHOMME, Yves (1966) : « Terme scientifique, mot linguistique, symbole scientifique », *Études de linguistique appliquée*, 1re série, 4, Paris, Didier, pp. 3-27.

GENTILHOMME, Yves (1982a) : « De la notion de notion à la notion de concept. Processus dynamique itératif d'acquisition des notions. Conséquences lexicales et didactiques », *T.C.R.S. (Travaux du Centre de Recherches Sémiologiques)*, 42, Université de Neuchâtel, pp. 66-89.

GENTILHOMME, Yves (1982b) : « Lecture d'un texte scientifique. Introduction », *Pratiques*, n° 35, oct., pp. 100-115.

GENTILHOMME, Yves (1984) : « Les faces cachées du discours scientifique. Réponse à Jean Peytard », *Langue Française*, 64, pp. 29-37.

GENTILHOMME, Yves (1985) : *Essai d'approche microsystémique. Théorie et pratique. Application dans le domaine des Sciences du langage*, Berne, Francfort/Main, New York, Peter Lang, 294 p.

GENTILHOMME, Yves (1992a) : « L'éclatement du signifié dans les discours technoscientifiques », *Signifiant, Référent, Réel*, CRsLE, vol. 20, Annales littéraires de l'Université de Besançon, diff. Les Belles Lettres, Paris, pp. 29-60.

GENTILHOMME, Yves (1992b) : « Initiation pédagogique au D.E.C. », *Études de linguistique appliquée, Hommage à Bernard Quemada, « dictionnairique et dictionnaires »*, pp. 161-174.

GENTILHOMME, Yves (1993) : « Réflexions sur les discours scientifiques », *Fachsprachentheorie Bd1 Fachsprachliche Terminologie, Begriffs- und Sachsysteme*, Methodologie, betreut und heraus gegeben von Theo Bungarten, Attikon Verlag, Tostedt, pp. 430-494.

GOUADEC, Daniel (1990) : *Terminologie. Constitution des données*, Paris, AFNOR.

GOUGENHEIM, Georges (1962) : *Dictionnaire fondamental de la langue française*, Paris, Didier, nouvelle édition revue et corrigée, (1re éd. 1958), 283 p.

GOUGENHEIM, Georges, RIVENC, Paul, MICHEA, René, SAUVAGEOT, Aurélien, (1964) : *L'Élaboration du français fondamental*, Paris, Didier, 1re éd. 1956, 302 p.

KOCOUREK, Rostislav (1982) : *La langue française de la technique et de la science*, 2e édition, Wiesbaden, Oscar Brandstetten, 259 p.

LACOMBE, Daniel (1964) : « Sur les mots et les symboles », APMEP, n° 239, pp. 343-355.

LE LIONNAIS, François (1979) : *Dictionnaire des Mathématiques*, A. BOUVIER et M. GEORGE (dir), Paris, PUF, 832 p.

LYONS, John (1970) : *Linguistique générale*, éd. angl. 1968, Paris, Larousse, 384 p.

MARTIN, Robert (1976) : *Inférence, antonymie et paraphrase. Éléments pour une théorie sémantique*, Paris, Klincksieck, 176 p.

MEL'ČUK, Igor A. et coll. (1984) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I*, 172 p., *II*, 1988, 332 p., et *III*, 1992, 323 p., Montréal, Les Presses de l'Université de Montréal.

MITTERAND, Henri (1992) : *Les mots français*, 8e édition, Paris, PUF, « Que sais-je ? », n° 270.

PEIRCE, Charles Sanders (1979) : *Écrits sur le signe*, G. Deladalle (dir.), éd. angl. 1931-1935, Paris, Le Seuil.

PHAL, André et Lucette BEIS (1971) : *Vocabulaire général d'orientation scientifique*, Paris, CREDIF, Didier, 128 p.

POTTIER, Bernard (1974) : *Linguistique générale, théorie et description*, Paris, Klincksieck, 340 p.

PUTNAM, Hilary (1988) : *Représentation et réalité*, trad. franç. 1990, Paris, Gallimard, 226 p.

QUATREMER, R. et J.-P. TROTIGNON (1981) : *Précis. Unités et grandeurs*, AFNOR, Nathan, 3e éd., 54 p.

SARMANT, Jean-Pierre (1978) : *Dictionnaire de physique*, « Faire le point », Paris, Hachette, 1850 articles, 288 p.

SAUSSURE, Ferdinand de (1915) : *Cours de Linguistique générale*, publié par Charles Bally et Albert Sechehaye avec la collaboration de Albert Riedlinger, Paris, Payot, 331 p.

SIBLOT, Paul (1993) : « Praxis et organisation du sens », *Encyclopédie et dictionnaires français (Problèmes de norme(s) et de nomenclature)*, par Daniel Baggioni (dir.) et le Centre Dumarsais, Univ. de Provence, pp. 49-57.

Spécial π, Supplément au Petit Archimède, mai 1980, éd. de l'ADCS, Amiens, 289 p.

STRAWSON, P. P. (1964) : « Phrase et acte de parole », trad. franç., *Langages*, mars 1970, n° 17, pp. 19-32.

TESNIERE, Lucien (1959) : *Éléments de syntaxe structurale,* 1re éd., (préface de Jean Fourquet), Paris, Klincksieck, 700 p.

TOURNIER, Jean (1991) : *Structures lexicales de l'anglais. Guide alphabétique,* Paris, Nathan, 190 p.

VUILLEUMIER, Viviane (1986) : *Signes et discours dans l'éducation et la vulgarisation scientifiques,* synthèse réalisée à partir des Actes des Sixièmes Journées Internationales sur l'Éducation Scientifique, Paris, Genève, Ministère de la Recherche (DIXIT) et de l'APDRS, 112 p.

27

Les relations notionnelles expérimentées dans les microglossaires de TERMISTI : du foisonnement à la régularité

Marc VAN CAMPENHOUDT

Institut supérieur de traducteurs et interprètes (ISTI), Bruxelles, Belgique

• *Abstract* •

The TERMISTI software was developed by the Institute of translators and interpreters (ISTI, Brussels) and serves to describe advanced terminologies within the framework of research into conceptual networks.

Several microglossaries, including a description of conceptual links, have been created. Today, this thesaurus offers the opportunity to check the hypotheses for a "relations grammar" and for the improvement of definitions. The paper tries to classify the links created by the various users of Termisti and to describe the limits of the predicate system particularly in order to control the number of links. This offers new prospects for a more user-friendly software which would meet the present-day publishing requirements.

Introduction

Lors des Deuxièmes Journées scientifiques du réseau Lexicologie, Terminologie et Traduction, à l'Université de Mons, le groupe de recherche TERMISTI avait présenté la toute première version de son logiciel de gestion de microglossaires terminologiques. Daniel Blampain (1991) avait alors insisté sur l'idéal théorique de nos travaux : fournir une information fiable et très spécialisée qui échapperait aux défauts imputables à l'importante luxuriance des grandes banques de données terminologiques. Il s'agissait de fournir une information sur microordinateur qui répondît aux besoins de traducteurs confrontés à des domaines de haute spécialité.

L'approche retenue se voulait notionnelle, dans la tradition viennoise, et se proposait de permettre la gestion des liens sémantiques qui lient les notions d'un même microdomaine. L'idéal poursuivi était de dégager des règles de gestion de ces liens et de proposer une nouvelle démarche d'exploitation du réseau notionnel.

Le fonctionnement du logiciel a été décrit en divers endroits (voir les références), aussi nous contenterons-nous de rappeler ici que les fiches suivent le modèle d'Eurodicautom, que leur organisation en base de données relationnelle permet d'intégrer une infinité de langues et de synonymes et que les liens entre les notions sont exprimés sous la forme d'une double prédication du type *A se compose de B* et *B est une partie de A* (tableau 1).

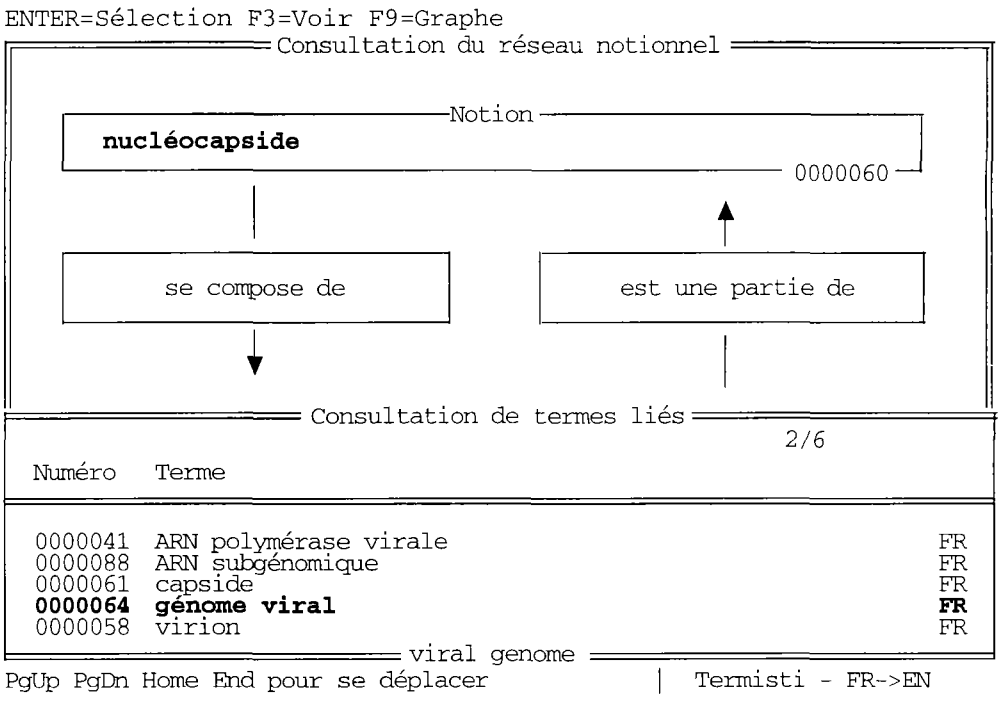


TABLEAU 1

Jusqu'à ce jour, notre but n'a pas été de réaliser un produit commercial. Il s'agissait avant tout de valider des hypothèses théoriques et d'appréhender les difficultés pratiques auxquelles se heurtent les terminologues qui entendent produire des microglossaires notionnels.

Au cours d'un projet étalé sur deux années, une chercheuse de notre unité, Pascaline Merten, a été chargée de réaliser des microglossaires en collaboration avec les

spécialistes de divers centres de recherches universitaires[1]. Parallèlement, trois étudiantes en fin de deuxième cycle en traduction ont réalisé des mémoires de fin d'études en collaboration avec l'Office de la langue française. Le sujet d'étude, la mécatronique automobile, a lui aussi débouché sur la création de microglossaires[2] TERMISTI.

Aujourd'hui, il est possible d'examiner chacune de ces microbases et de tirer les premières conclusions de l'expérience. Plus particulièrement, nous avons choisi de traiter ici de l'un des points les plus originaux du logiciel : la gestion des liens notionnels.

Les liens notionnels

Nettement inspiré des théories viennoises et de la norme ISO 704 (1987), le logiciel permet au départ d'établir deux grandes catégories de liens : les liens hiérarchiques et les liens coordonnés. Pour avoir activement participé à la phase d'analyse qui a précédé l'élaboration du logiciel, nous pouvons expliquer ce choix par l'importance que nous accordions alors à cette distinction qui fonde l'essentiel de la théorie des notions. Parmi les liens hiérarchiques, nous prévoyions bien entendu de ranger les relations espèce-genre et partie-tout, alors que les liens coordonnés devaient permettre d'exprimer un grand nombre de relations sémantiques, notamment celles liées aux dimensions spatiale et temporelle.

Nous avions prévu de permettre à l'utilisateur du logiciel de créer une infinité de liens relevant de chacune de ces catégories grâce à un système de double prédication. En dehors de l'impératif de classement hiérarchique/coordonné, le terminologue utilisant le logiciel disposait donc d'une totale liberté de choix dans la manière d'exprimer le lien entre les notions. En effet, nous avions décidé, pour cette phase exploratoire, de ne rattacher les prédicats qu'au seul microglossaire concerné, quitte à réintroduire une forme de luxuriance là où notre idéal attendait rigueur et clarté. Cette grande liberté devait permettre d'analyser, dans un second temps, l'usage de chacun et d'en déduire des points de convergence qui fonderaient une théorie du réseau notionnel. Confrontés à des vocabulaires de spécialité fonctionnant dans toute la cohérence d'un sous-domaine homogène, les collaborateurs de TERMISTI espéraient découvrir des phénomènes plus complexes que ceux décrits par divers manuels de terminologie à partir d'exemples inévitablement disparates et fort généraux.

Avantages et limites du système de prédication

Nous avons déjà eu l'occasion de décrire en d'autres lieux les avantages d'un système qui exprime chaque lien notionnel sous la forme d'un prédicat. Outre qu'il supplée à l'absence de définition pour les notions ultraspécialisées situées au bas de l'arborescence, il permet de traduire une même relation sémantique dans une infinité de langues et se révèle très formateur pour le traducteur qui s'initie au domaine. Pourtant,

1. - Phytovirologie : 206 notions - 630 termes EN/FR (P. Merten).
- Science du sol : 91 notions - 216 termes EN/FR (P. Merten).
- Télédétection aérospatiale : 253 notions - 558 termes EN/FR (P. Merten).
- Effets de serre : 50 notions - 185 termes EN/FR (P. Merten).
2. - Suspension (mécatronique automobile) : 130 notions - 626 termes EN/FR (I. Simal).
- Freinage (mécatronique automobile) : 109 notions - 681 termes EN/FR (I. Libert).

la liberté de prédication offerte au terminologue qui conçoit un microglossaire peut paradoxalement engendrer divers désordres dans la communication.

Expression du prédicat

Le premier obstacle auquel on songe est lié à l'essence même des langues de spécialité, lesquelles peuvent exprimer un lien notionnel à travers des prédicats relativement figés, car propres à la terminologie décrite. Force est de constater que ce phénomène est absent des domaines traités, dans la mesure où l'on n'a pas eu affaire à des langues de spécialité de vieille tradition qui auraient développé les catégories non nominales ou des expressions idiomatiques. De plus, on peut penser que les terminologues qui ont créé les réseaux ont veillé à utiliser des prédicats aisément compréhensibles.

Un problème plus tangible est lié à l'expression de la relation partie-tout, laquelle apparaît très difficile à manier, comme l'ont déjà montré Winston *et al.* (1987)[3]. Ainsi, dans la base consacrée à la pédologie, on trouve d'une part la relation *est une partie de* vs *se divise en* et d'autre part la relation *est une partie de* vs *se compose de.*

sol ---- *se divise en* ----> agrégat élémentaire (pédologie)
agrégat élémentaire ---- *est une partie de* ----> sol

granulométrie ---- *se compose de* ----> fraction minérale (pédologie)
fraction minérale ---- *est une partie de* ----> granulométrie

On peut bien sûr songer qu'il ne s'agit là que d'un problème de formulation d'une nuance, qui pourrait être résolu par un effort d'imagination. Dans le microglossaire consacré à la phytovirologie, on identifie d'ailleurs deux relations très proches, mais qui s'expriment d'une manière plus particulière :

Unité fonctionnelle : *est l'unité fonctionnelle de* vs *se divise en* (phytovirologie)
gène ---- *est l'unité fonctionnelle de* ----> chromosome
Unité fonctionnelle : *est l'unité structurelle de* vs *est constitué de* (phytovirologie)
bâtonnet rigide ---- *est l'unité structurelle de* ----> virus à particules rigides

À la lumière du domaine traité, il est effectivement apparu qu'une distinction théorique permettait de mieux discerner les deux relations : un tout peut être constitué de parties répétitives et identiques qui forment sa structure ou d'un assemblage de parties répétitives qui ont chacune leur fonction. Il reste que l'on demeure confronté à l'opposition entre la langue du profane, imprécise mais très compréhensible, et la langue du spécialiste, qui suppose une initiation.

Doublon

Un logiciel exige d'autant plus de rigueur dans son usage qu'il est ouvert à la créativité. Si le terminologue utilise tantôt le prédicat *A est un type de B,* tantôt le prédicat

3. Nous décortiquons ces théories dans une thèse que nous préparons à l'Université Paris-Nord sous la direction du professeur Pierre Lerat.

A est un B, il aboutit à créer deux liens différents qui rendent compte d'une seule et même relation sémantique. Toutes les relations peuvent être concernées par ce type de doublon, généralement identifiable par l'usage d'un même prédicat inverse.

Phénomène : *peut apparaître sur* vs *peut entraîner* (freinage)
couple de lacet ---- *peut apparaître sur* ----> sol à coefficient d'adhérence asymétrique
sol à coefficient d'adhérence asymétrique ---- *peut entraîner* ----> couple de lacet
Phénomène : *peut apparaître en cas de* vs *peut entraîner* (freinage)
blocage de roue ---- *peut apparaître en cas de* ----> sol à coefficient d'adhérence asymétrique
sol à coefficient d'adhérence asymétrique ---- *peut entraîner* ----> blocage de roue

Le cas peut s'expliquer par un manque de rigueur comme par la présence de deux terminologues aux goûts linguistiques divergents. À cet égard, la gestion des liens se heurte aux problèmes classiques de cohérence propre à n'importe quelle base de données.

Prédicat et « multidimensionnalité »

Le concept de *multidimensionnalité,* actuellement étudié par Ingrid Meyer (École de traducteurs et interprètes de l'Université d'Ottawa) et Lynne Bowker (University of Manchester Institute of Science and Technology) constitue la source de bien des difficultés. Une même notion peut effectivement être décrite sous plusieurs aspects. Le problème est très facilement résolu si l'on se contente de fournir des définitions, puisque celles-ci peuvent adopter systématiquement tel ou tel point de vue. Toutefois, lorsqu'on établit un réseau, tous les liens sont envisageables, ce qui peut notamment créer des problèmes de double appartenance. Déjà Felber (1987 : 101) montre que la notion *vol affrété transatlantique* peut être vue comme hyponyme de *vol affrété* et comme hyponyme de *vol transatlantique*. De même, dans la relation partie-tout, il n'est pas rare d'observer que tel composant appartient à la fois à deux dispositifs différents.

Parallèlement, le caractère facultatif ou potentiel d'un certain nombre de relations a visiblement amené le concepteur du réseau notionnel du freinage (mécatronique automobile) à opter pour des formulations adaptées à une certaine souplesse de vue :

Composant : *se compose de* vs *est une partie de* (suspension)
suspension électronique ---- *se compose de* ----> pompe
Composant facultatif : *peut se composer de* vs *peut être une composante de* (suspension)
suspension électronique ---- *peut se composer de* ----> amortisseur adaptatif
Résultat : *a pour résultat* vs *est le résultat de* (freinage)
système de freinage ---- *a pour résultat* ----> couple de freinage
Résultat possible : *peut avoir pour résultat* vs *peut résulter de* (freinage)
suspension électronique ---- *se compose de* ----> antiplongée

Ce type de formulation est à tout le moins ambigu : le dernier prédicat signifie-t-

il que la relation est facultative ou que la suspension électronique peut engendrer d'autres effets que l'antiplongée ?

Le caractère facultatif ou potentiel apparaît fréquemment et risque de se confondre avec l'expression de la multidimensionnalité. En effet, un prédicat ne permet pas d'exprimer aisément la nuance séparant ces cas et nous nous demandons si un autre système de distinction, fondé par exemple sur un signal visuel, ne serait pas préférable à celui de l'expression linguistique.

Instabilité du prédicat

Ces différentes observations montrent qu'à tout le moins, il conviendrait de revoir le mode d'identification des liens dans le logiciel TERMISTI. Chaque prédicat devrait s'accompagner d'un code d'identification de la relation, laquelle serait préalablement décrite dans une « bibliothèque des liens ». La formulation du lien pourrait être retravaillée à tout moment, de sorte qu'une certaine harmonie règne entre les divers microglossaires et que l'on puisse ultérieurement opérer un retour analytique sur l'usage. Lorsqu'un logiciel est à ce point sous-tendu par des hypothèses théoriques, il convient de veiller à ce que l'utilisateur retrouve une même unité de pensée dans les différents microglossaires consultés.

Distinction entre prédicat définitionnel et lien structurant le réseau : une piste face à la multidimensionnalité ?

À propos des liens hiérarchiques et coordonnés

Nous sommes convaincu qu'une piste de recherche très intéressante consiste à distinguer d'une part les liens qui contribuent à l'ossature du réseau et d'autre part les liens qui fournissent une indication sémantique plus ponctuelle pour situer telle notion par rapport à telle autre. On sait en effet que nombre de définitions se fondent sur la relation hyponymique et éventuellement sur la relation méronymique (partie-tout[4]). Les autres relations varient bien davantage en fonction des sujets abordés.

L'idée d'une distinction entre liens hiérarchiques et coordonnés nous semble pourtant appropriée, du moins si on la reprécise à la lumière des apports de la sémantique lexicale et de la psychologie cognitive. Ainsi, Cruse (1986 :181-196) propose de distinguer des hiérarchies arborescentes (*branching hierarchies*) et des hiérarchies non arborescentes (*non-branching hierarchies*). Les premières s'identifient *grosso modo* aux relations hyponymiques (espèce-genre) et méronymiques (partie-tout) ; les secondes unissent par une même relation des suites de notions selon des modèles structuraux bien précis qu'il dénomme notamment *chaîne, cycle* et *hélice*. Très clairement, il apparaît que les notions qui forment une hiérarchie non arborescente sont généralement liées par un lien de cohyponymie ou de coméronymie.

4. Conformément à une habitude propre à notre équipe de recherche, nous adoptons le terme méronymie pour désigner les différentes relations partie-tout. Le *méronyme* est la notion subordonnée et l'*holonyme* la notion superordonnée. Nous évitons toute appellation qui, comme *hyperonyme partitif*, fait référence à la relation espèce-genre.

Nous sommes tenté de ne définir comme *coordonnées* que celles des relations qui unissent des notions par un lien de cohyponymie ou de coméronymie. À notre sens, si l'on veut que le réseau demeure lisible, il convient de préciser quelles sont les relations qui, par rapport à telle ou telle arborescence, structurent les cohyponymes ou coméronymes dans un ordre pertinent. Ainsi, selon le caractère différenciateur activé, telle ou telle relation non hiérarchique peut servir à structurer le lien de cohyponymie.

Les relations fonctionnelles

Que deviennent, dans le cadre d'une telle vision, les autres relations qui forment l'armée des relations traditionnellement dites *coordonnées* ou *non hiérarchiques* ? Bien entendu, elles demeurent pertinentes car sémantiquement nécessaires, mais elles ne fondent pas à notre sens la structure fondamentale du réseau notionnel. Il s'agit de prédicats présents dans la définition de chaque notion et qui peuvent prendre des aspects très différents selon le domaine concerné. Les pères de la terminologie ont surtout retenu des relations spatio-temporelles et de cause-effet. Aujourd'hui, les chercheurs insistent sur l'importance des relations dites *fonctionnelles.*

Une nature pseudo-hiérarchique

Dans un article paru en 1990, Pierre Lerat proposait de considérer sous le nom de *relations fonctionnelles* les liens sémantiques permettant de situer les notions entre elles tout en fournissant leurs propriétés essentielles. Celles-ci sont, en effet, aussi informatives que les traditionnelles mentions de l'hyperonyme et du méronyme. L'auteur proposait d'assimiler ce type de prédicat à une forme de relation hiérarchique et utilisait d'ailleurs l'expression *hyperonymie fonctionnelle.*

Une première taxonomie des relations notionnelles utilisées dans nos microglossaires a confirmé l'intérêt de ces innombrables relations de type « cas sémantique ». Pascaline Merten (1992) a également proposé de considérer ces relations dites *fonctionnelles* comme hiérarchiques, à la différence des relations spatio-temporelles, cataloguées comme coordonnées, au sens classique du terme.

Adjuvant :	*est l'adjuvant de*	vs	*a pour adjuvant*
Agent :	*est l'agent de*	vs	*a pour agent*
Destinataire :	*est le destinataire de*	vs	*a pour destinataire*
Temps/Durée :	*est la durée de*	vs	*dure*
Instrument :	*est l'instrument de*	vs	*a pour instrument*
Résultat :	*est le résultat de*	vs	*a pour résultat*
etc.			

Personnellement, nous éprouvons quelques difficultés à utiliser dans ce cadre l'appellation *hiérarchique* : à nos yeux, celle-ci mérite de se limiter – comme le propose la norme ISO 704 (1987 : 3-4) – aux cas de l'inclusion logique (est inclus dans l'ensemble des X) et de l'inclusion ontologique (est présent dans X), seules capables de fonder des arborescences sur plusieurs niveaux. Toute autre relation peut certes être appréhendée comme hiérarchique, mais il ne s'agira jamais que d'une vue de l'esprit qui perçoit une forme de primauté ou de dépendance conceptuelle dans des rela-

tions où A est perçu comme « plus fort » que B, car il est *la cause de* B, *l'agent de* B, *le support de* B, etc.[5]. On notera d'ailleurs que lorsqu'il parle de hiérarchie non arborescente, Cruse (1986) ne fait que considérer des cas où une relation semblable unit plusieurs notions en un modèle structurant. Nous proposons donc de définir comme *coordonnée*, c'est-à-dire comme utile à la structure fondamentale du réseau, toute relation qui fonde ponctuellement une distinction entre co-hyponymes ou co-méronymes.

Un rôle définitionnel

Les autres relations fonctionnelles, spatiales, temporelles ou autres, moins utiles à visualiser sous forme de graphe – elles sont valables pour tout hyponyme –, jouent néanmoins un rôle définitionnel très important. À ce titre, elles méritent d'être décrites à travers des prédicats appropriés. Il est évident, lorsqu'on consulte les réseaux notionnels créés, que les relations fonctionnelles ont pris une ampleur inattendue de même qu'il a fallu affiner l'expression de la relation partie-tout. Cet état de fait est dû, sans conteste, à la nécessité de fournir une information sémantique pertinente en l'absence de définition attestée.

La terminologie ne pouvant se contenter du flou propre à de nombreux réseaux sémantiques, la typologie des relations fonctionnelles, ou tout au moins leur formulation, apparaît fort liée au domaine considéré. En mécatronique automobile, par exemple, on a utilisé des relations comme *émet, utilise, contrôle, calcule, reçoit, stocke* etc. À tout le moins, les théories initiales se trouvent bouleversées par des prédicats comme *est une condition d'intervention de, est caractérisé par la présence de, transmet le liquide à, est contrecarré par, est effectué lors de, est la zone de travail de* etc.

Derrière la luxuriance, la régularité ?

Dans la mesure où ils ont été créés, de tels prédicats ont assurément été jugés nécessaires à la fourniture d'une information correcte au traducteur. Certes, il est bon de constater que le système est suffisamment ouvert pour permettre à chacun d'établir les relations qui lui paraissent pertinentes. Toutefois, il conviendrait, dans un deuxième temps, de mettre le réseau des relations fonctionnelles à plat et d'y rechercher des formes de régularité. Force est de constater que, jusqu'à présent, on s'est plus attaché à décrire les structures engendrées par les relations hyponymiques que celles qui naissent de la prise en compte des relations méronymiques et des relations fonctionnelles.

Les travaux de TERMISTI ont permis de dresser une première typologie des relations fonctionnelles. Notre hypothèse est que derrière les prédicats utilisés se trouvent parfois des relations proches ou identiques. Ainsi, peut-on voir des variantes de la relation d'adjuvant derrière des expressions comme *contrôle, calcule, contrecarre* etc. Les relations fonctionnelles pourraient donc être, elles-mêmes, perçues comme incluses au sein d'une hiérarchie espèce-genre : certaines seraient en réalité des types plus particuliers et exigeraient dès lors des formulations plus affinées.

5. Nous étendons cette critique à la relation de caractérisation présentée comme hiérarchique par P. Merten (1992). Ce lien d'attribution d'une propriété se comporte comme de nombreuses relations fonctionnelles et se transfère aux éventuels hyponymes par la loi d'héritage.

Comme c'est le cas pour la relation méronymique, il est sans doute possible de préciser à l'aide de traits distinctifs quelles sont les caractéristiques minimales de telle ou telle relation. Winston *et al.* (1987) proposent ainsi de distinguer différents types de relations partie-tout sur la base de grilles sémiques aujourd'hui expérimentées à travers nos microglossaires. Nous sommes personnellement porté à croire que la validation de traits sémiques permet parfois de déterminer avec plus de clarté quelle relation est en jeu. Même si tout système componentiel présente l'inconvénient de la lourdeur, la multiplication des traits permet un affinement considérable de l'information. Ainsi, la traditionnelle relation composant-objet possède notamment deux variantes hyponymiques, selon que le composant est une unité structurelle ou une unité fonctionnelle. Dans le premier cas, l'objet est formé de l'assemblage répétitif de pièces identiques qui possèdent toutes le même rôle, dans le second cas, chacune de ces pièces peut être appelée à jouer un rôle particulier et possède des caractéristiques qui lui sont liées (voir le paragraphe *Expression du prédicat*).

Jusqu'à présent, nous avons identifié une dizaine de traits qui permettent de distinguer le type de méronymie[6]. Bien sûr, nous ne pensons pas qu'il soit systématiquement nécessaire d'aller aussi loin dans la distinction des relations. Chaque discipline possède ses exigences en la matière et il ne convient pas d'aller au delà des distinctions qui paraissent nécessaires aux spécialistes du domaine. Ce qui est vrai des relations partie-tout l'est également des relations fonctionnelles, lesquelles semblent fort liées au domaine. Dans les microglossaires de mécatronique on trouve la variété des prédicats déjà cités alors que celui de phytovirologie se borne à utiliser les prédicats présentés par Pascaline Merten (1992) à TAMA'92.

Gérer le réseau à partir de modèles organisationnels

Des figures hautement significatives

Le logiciel TERMISTI inclut une fonction « graphe » très élémentaire qui représente le réseau notionnel sans préciser la valeur des liens qui unissent les notions. Cette fonction connaît un vif succès auprès des utilisateurs du logiciel, mais s'avère rapidement inutile à mesure que croît le nombre de relations. En outre, le mode de figuration ne permet qu'un ordonnancement strictement vertical qui ne reflète en rien les modèles structuraux déjà évoqués[7].

À nos yeux, l'avenir appartient à un système qui permettrait de sélectionner les relations souhaitées, d'isoler certaines notions et surtout de faire ressortir les structures jugées pertinentes. Par exemple, il devrait être possible de demander au logiciel d'identifier les parties du réseau où des notions cohyponymes sont reliées par des modèles d'agencement identifiables.

La possibilité de faire figurer une image en trois dimensions en sélectionnant les parties du réseau jugées intéressantes devrait permettre, quant à elle, de beaucoup mieux faire ressortir les liens partie-tout qui unissent plusieurs arborescences espèce-genre

6. On citera par exemple le caractère comptable, le caractère séparable, la fonction particulière, l'emplacement prédéterminé, le matériau homogène, l'appartenance à une même typologie etc.
7. De ce point de vue, le projet *Code* développé par Ingrid Meyer semble particulièrement en pointe, même si les exemples de graphes paraissent eux aussi souvent touffus.

(tableau 2). Faute de cette possibilité, le bruit généré par le foisonnement des relations empêchera de réellement comprendre intuitivement le fonctionnement d'un sous-domaine.

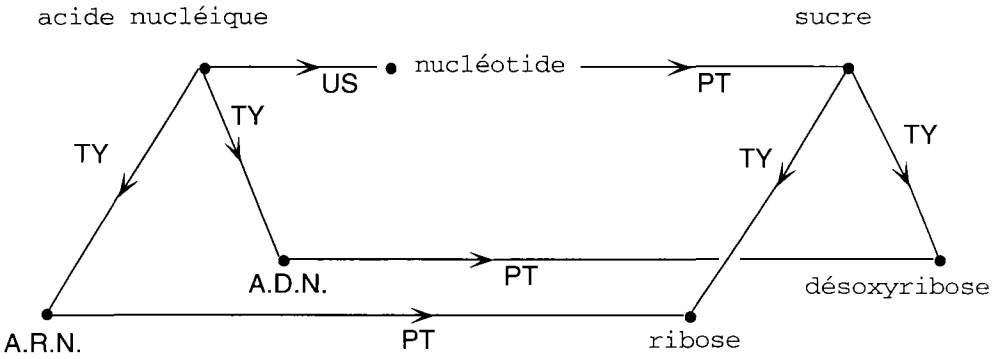


TABLEAU 2 : Tentative de représentation du problème posé par Merten (1992 : 220).

Il reste qu'il faut conserver en mémoire que le réseau ne constitue pas une représentation de la réalité, fût-elle abstraite. Il n'est que la notation symbolique de liens sémantiques entre les notions, comme nous le rappellera un exemple emprunté à l'astronomie. Il est aisé de reproduire schématiquement la succession des phases de la lune au cours du mois lunaire et d'associer dans le même dessin les périodes de quadrature et de syzygie qui se succèdent par deux fois au cours dudit mois (tableau 3).

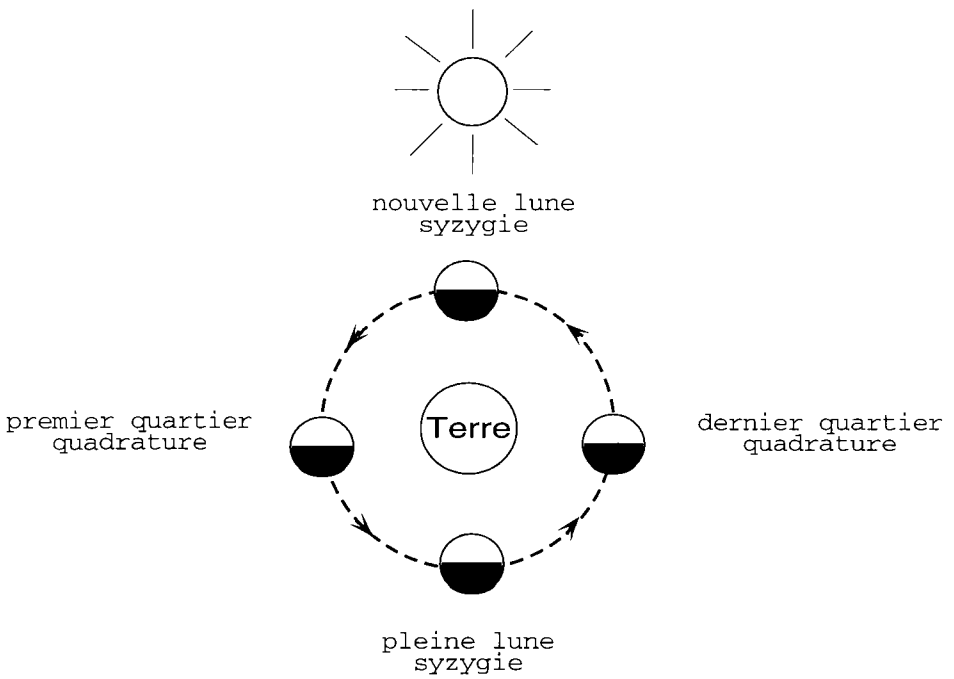


TABLEAU 3.

Dans un réseau notionnel, cela s'avère impossible, à moins de faire figurer deux fois les notions de quadrature et de syzygie (tableau 4). À cet égard, il serait vain de demander au réseau de jouer le rôle d'une illustration. On devrait, par contre, s'attacher à mieux étudier l'interaction entre ces deux modes de représentation et la manière dont ils se complètent.

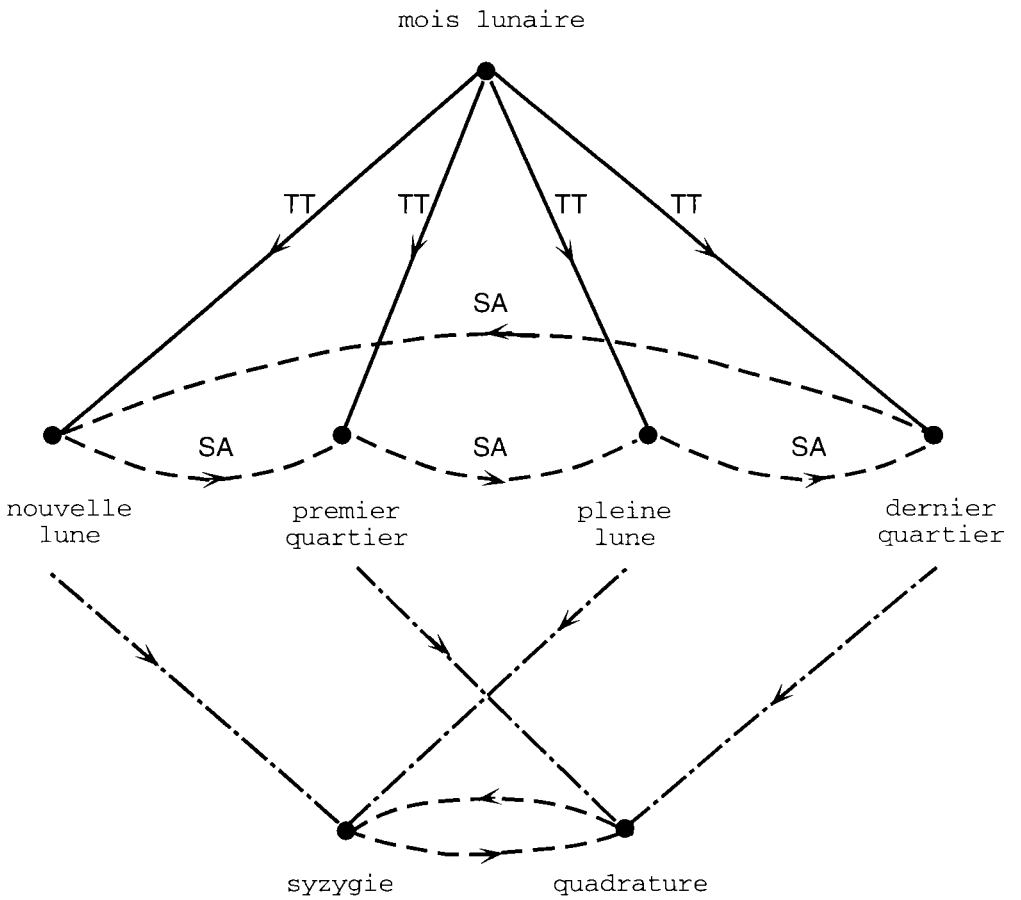


TABLEAU 4.

Une aide à la gestion du réseau

L'expérience nous enseigne qu'il est difficile de se montrer systématique dans la construction du réseau. La fonction « graphe » permet déjà de visualiser le travail du terminologue et de dépister les éventuels oublis. Sans songer à un système d'interprétation des contextes et définitions, on peut raisonnablement envisager de créer des

algorithmes qui aideraient le terminologue à élaborer son réseau sur la base de quelques règles liées aux modèles structuraux identifiés. En effet, un modèle implique une certaine régularité et donc la possibilité d'envisager des règles de gestion.

Nul n'ignore que la relation espèce-genre se prête à la transitivité et implique un héritage des caractéristiques de l'hyperonyme. La même transitivité est souvent décrite comme s'appliquant, elle aussi, à la relation partie-tout. Pourtant, l'expérience nous a appris à distinguer des variétés de relations méronymiques qui ne se prêtent guère à la transitivité dès lors qu'elles s'expriment à travers une prédication qui les mélange[8]. Il convient donc de se méfier des règles trop systématiques ; si l'avenir devait nous permettre d'implanter des fonctions de gestion du réseau, nous veillerions à les concevoir comme des aides à la décision plutôt que comme des outils de création automatique des liens.

Une meilleure connaissance des modèles structuraux qui sous-tendent le réseau devrait ainsi permettre de concevoir un programme qui, sur la base des relations déjà établies, proposerait de compléter le réseau en suivant le modèle établi. Par exemple, la relation TT *être une tranche temporelle de,* forme de relation méronymique illustrée par le réseau du mois lunaire, suppose que les coméronymes soient tous liés par une même relation coordonnée temporelle *succède à* (relation SA). Plusieurs règles logiques en découlent :

● Constatant que plusieurs notions sont des phases d'un même holonyme TT, le programme pourrait proposer de les lier par un lien coordonné temporel.

X, Z (méronyme TT *est une phase de*)
Y, Z (méronyme TT *est une phase de*)
=> *Validez, s'il y a lieu, le lien coordonné temporel adéquat* :
– X, Y (*succède à*)
– Y, X (*succède à*)

● Inversement, si plusieurs notions sont liées par un même lien de succession temporelle formant une chaîne homogène, un cycle ou une hélice et que l'une a été reliée par une relation méronymique TT avec une notion englobante, le programme pourrait proposer d'établir un même lien méronymique TT entre la notion englobante et les notions incluses dans le modèle susmentionné.

X, Y (*succède à*)
X, Z (méronyme TT *est une phase de*)
=> *Validez, s'il y a lieu, la proposition suivante* :
– Y, Z (méronyme TT *est une phase de*)

● Le système pourrait même permettre d'affiner la relation envisagée. Si dans le cas précédent, le lien méronymique n'est pas défini comme temporel, le programme pourrait en faire la proposition du fait de la présence d'un chaînage temporel.

8. Par exemple, la transitivité des parties est difficile à exprimer lorsqu'on mélange la matière et le composant : si le pare-choc de la voiture est en polyéthylène, l'on ne peut pas dire que la voiture est en polyéthylène, même s'il est vrai que le polyéthylène fait partie de la voiture.

X, Y (*succède à*)
X, Z (méronyme PT *est une partie de*)
Y, Z (méronyme PT *est une partie de*)
=> *Validez, s'il y a lieu, le lien coordonné temporel adéquat* :
- X, Z (méronyme TT *est une phase de*)
- Y, Z (méronyme TT *est une phase de*)

Ceci ne constitue bien sûr qu'un exemple parmi de nombreuses règles que nous avons déjà envisagées et que nous espérons pouvoir un jour développer dans un nouveau logiciel.

Fonder l'ordonnancement macrostructurel sur le réseau notionnel

L'un des principaux apports du réseau devrait être de permettre un ordonnancement véritablement logique des notions dans les publications terminographiques. Jusqu'à présent, en effet, la plupart des dictionnaires de traduction se fondent soit sur l'ordre alphabétique des termes d'une langue (l'anglais en règle générale), soit sur un ordre dit « logique » qui tente de raccrocher les notions entre elles, sans passer par le critère d'une langue. Nos études montrent que cet ordonnancement est fréquemment basé sur des contiguïtés mentales qui, lorsqu'elles ne varient pas selon l'appréhension de chaque spécialiste, échappent du moins à la compréhension du néophyte.

L'idée maîtresse serait de fonder l'ordre logique sur les relations notionnelles inscrites dans la base de données. Un tel ordonnancement des notions présenterait de réelles vertus didactiques pour le traducteur qui tente de comprendre l'organisation d'un domaine. Ainsi, s'agissant de décrire les notions propres à un assemblage particulier, on pourrait sélectionner les quatre relations suivantes et décider de suivre un algorithme précis.

TY = relation espèce-genre
PT = relation partie-tout
DD = relation devant-derrière
HT = relation au-dessus–en dessous

Règle : suivre la relation TY (versant X *est un* Y) en mentionnant pour chaque hyponyme les notions liées par la relation PT (versant X *est une partie de* Y), ordonnancer les cohyponymes et les coméronymes en fonction de la relation DD (versant X *est devant* Y) ou HT (versant X *est au-dessus de* Y).

Hyperonyme TY
  Méronymes PT (classement des coméronymes en fonction de la relation DD ou HT)
    Cohyponyme TY n° 1 (classement des cohyponymes en fonction de la relation DD ou HT)
      Méronyme PT n° 1.1 (classement des coméronymes en fonction de la relation DD ou HT)
    Cohyponyme TY n° 2 (classement des cohyponymes en fonction de la relation DD ou HT)

Méronyme PT n° 2.2 (classement des coméronymes en fonction de la relation DD ou HT)

Ici encore, la mise en place de ces mécanismes de connaissance ne sera rendue possible que par l'étude approfondie des principes qui structurent les arborescences[9]. Il serait utopique de penser que n'importe quel microdomaine peut se prêter à une telle organisation, notamment parce que, nous l'avons vu, divers microdomaines se décrivent à travers des relations fonctionnelles plus qu'à travers des relations hiérarchiques classiques. Pourtant, les grandes figures d'organisation existent et à défaut de pouvoir les exploiter, on peut au moins envisager d'imprimer les notions dans un ordre pseudo-logique où les liens seraient clairement mentionnés entre les notions que l'auteur affirme présenter dans un ordre pertinent.

Conclusion

Les réseaux établis dans le cadre des recherches de TERMISTI ne semblent pas accréditer l'idée que des microdomaines très spécialisés appellent des liens notionnels eux-mêmes très spécifiques. Les technologies de pointe posent avant tout des problèmes définitoires que notre équipe a tenté de résoudre à l'aide de relations notionnelles qui soient accessibles au profane. À cet égard, la formulation du lien, et plus particulièrement l'expression de telle nuance propre à telle spécialité, constitue une source de confusion. L'idée d'un recours à des messages non verbaux et à la validation de traits distinctifs semble une perspective intéressante pour contourner cette difficulté.

Le caractère éclectique des microdomaines retenus a confirmé l'idée qu'il convient de dépasser la vision initiale de la terminologie, fondée sur l'étude de domaines à l'organisation particulièrement régulière (machinerie, aéronautique, construction navale etc.). La prise en compte de spécialités de pointe conduit à considérer toutes les notions particulières d'un sous-domaine et non plus les seules notions générales du domaine. Si l'on souhaite en établir le réseau notionnel, on envisagera non seulement les relations d'inclusion logique ou ontologique, mais aussi les relations dites *fonctionnelles,* dont l'apport sémantique paraît primordial.

Dès à présent, l'émergence de ce nouveau type de relation nous incite à fournir un nouvel effort théorique visant à décrire leurs modèles d'agencement par rapport aux relations hyponymiques et méronymiques. Une exploitation logique du réseau garantit, en effet, une meilleure cohérence de l'information et un progrès dans le transfert des connaissances.

Références

BLAMPAIN, D., PETRUSSA, Ph. et M. VAN CAMPENHOUDT (1991) : « À la recherche d'écosystèmes terminologiques », Clas, A. et Safar, H. (dir), *L'environnement traductionnel. La station de travail du traducteur de l'an 2001,* Actes des Deuxièmes Journées scientifiques du Réseau thématique de recherche « Lexicologie, terminologie et traduction », Mons, 25-27 avril 1991, Sillery et Montréal, Presses de l'Université du Québec et AUPELF-UREF, Universités francophones, actualité scientifique, pp. 273-282.

9. Nous pensons notamment au risque de bouclage engendré par un tel algorithme.

CHAFFIN, R. et D. J. HERRMANN (1988) : « The Nature of Semantic Relations: a Comparison of Two Approaches », Evens, M. W. (dir), *Relational Models of the Lexicon. Representing Knowledge in Semantic Networks*, Cambridge, New York, Cambridge University Press, Studies in natural language processing, pp. 288-334.

CHAFFIN, R., HERRMANN, D. J. et M. WINSTON (1988) : « An Empirical Taxonomy of Part-Whole Relations: Effects of Part-Whole Relation Type on Relation Identification », *Language and Cognitive Processes*, vol. 3, n° 1, pp. 17-48.

COSTERMANS, J. (1980) : *Psychologie du langage*, Bruxelles, Mardaga, Psychologie et sciences humaines.

CRUSE, D. A. (1979) : « On the Transitivity of the Part-Whole Relation », *Journal of Linguistics*, vol. 15, n° 1, pp. 29-38.

CRUSE, D. A. (1986) : *Lexical Semantics*, Cambridge, London, New York, etc., Cambridge Univeristy Press.

FELBER, H. (1987) : *Manuel de terminologie*, Vienne, Infoterm.

ISO 704 (1987) : *Principes et méthodes de la terminologie*, s.l., Organisation internationale de normalisation (ISO/TC 37).

ISO R 1149 (1969) : *Présentation des vocabulaires systématiques multilingues*, s.l., Organisation internationale de normalisation (ISO/TC 37).

LERAT, P. (1990) : « L'hyperonymie dans la structuration des terminologies », *Langage*, n° 98, pp. 79-86.

MERTEN, P. (1992) : « Apport des relations notionnelles à la description terminologique », *TAMA '92, Deuxième symposium TermNet: Applications terminologiques et microordinateurs*, 5-6 juin 92, Vienne, TermNet, pp. 201-228.

MERTEN, P. (1993) : *Élaboration de microglossaires informatisés pour les langues de spécialité*, Rapport final, 15 février 1993, Bruxelles, Institut supérieur de traducteurs et interprètes et Communauté française de Belgique (polycopié).

MERTEN, P., MERTENS, J. et M. VAN CAMPENHOUDT (1993) : « Microglossaire, réseau notionnel et gestion informatique. Une expérience de recherche en Communauté française de Belgique », Gouadec, D. (dir), *Terminologie & terminotique: outils, modèles & méthodes. Actes de la Première université d'automne en terminologie*, Rennes 2, 21-26 sept. 1992, Paris, La Maison du dictionnaire, pp. 277-293.

SAGER, J. C. (1990) : *A practical Course in Terminology Processing*, Amsterdam et Philadelphia, John Benjamins Publishing Company.

VAN CAMPENHOUDT, M. (1991) : « TI, le logiciel d'expérimentation notionnelle de Termisti », *Terminologies nouvelles*, n° 5, pp. 11-14.

WINSTON, M. E., CHAFFIN, R. et D. HERRMANN (1987) : « A Taxonomy of Part-Whole Relations », *Cognitive Science*, vol. 11, n° 4, pp. 417-444.

28

De la focalisation à l'amplification : nouvelles perspectives de représentation des données terminologiques

Ingrid MEYER et Bruce MCHAFFIE[1]

École de traduction et d'interprétation, Université d'Ottawa, Canada

• *Abstract* •

The purpose of this paper is to outline two research directions in term bank design that promise wider-angle views *of terminological data than does the conventional term bank. The first direction involves the knowledge base model deriving from Artificial Intelligence, and is based on an explicit modelling of the conceptual structures associated with a term. The second direction involves the electronic corpus and corpus analysis tools, which provide glimpses of emergent conceptual structures, as well as a rich source of evidence about the use of terms in context. We propose that these two directions – grounded in concept analysis and corpus analysis, respectively – are highly complementary. We also argue that they are compatible with traditional terminology work, which has always emphasized concept analysis and the use of specialized documentation. Our paper is illustrated with examples from a terminological knowledge base called COGNITERM, developed at the Artificial Intelligence Laboratory of the University of Ottawa, Canada.*

Les banques de terminologie traditionnelles et l'« effet de focalisation »

Dans un article publié en 1987, Barbara Moser-Mercer décrit certains facteurs qui réduisent grandement l'utilité des systèmes informatisés pour les traducteurs et les terminologues. Nous aborderons ici l'« effet de focalisation »[2], terme qui caractérise une

1. La présente communication a été adaptée de l'anglais par René Morin. Le texte original s'intitule « *From "Peepholes" to "Wide-Angle Views" : Exploring New Vistas in Terminological Data Representation* » (Working Paper 93-01, Laboratoire d'intelligence artificielle, Université d'Ottawa, Canada).
2. C'est ce que Barbara Moser-Mercer appelle le « *peephole effect* ».

situation dans laquelle un usager voudrait une vue d'ensemble de l'information mais n'obtient qu'une perspective réduite, comme celle que donne le judas d'une grande salle de bal. L'usager, frustré de n'avoir pu obtenir la perspective souhaitée, s'insurge alors contre le système utilisé (Moser-Mercer 1987 : 157).

Dans cet article, nous utiliserons ce concept de « perspective réduite » pour illustrer certaines contraintes associées au modèle actuel de représentation des données terminologiques, c'est-à-dire la banque de terminologie traditionnelle. Dans ce modèle, l'information terminologique est présentée sous une perspective bien connue des traducteurs et autres spécialistes de la langue : la fiche terminologique. Comme le montre la figure 1, les champs-clés que comportent habituellement[3] les fiches extraites des banques de terminologie multilingues sont la désignation de la notion dans deux langues ou plus, l'indication du domaine, la définition et les contextes d'utilisation.

SUBJECT FIELD optical discs	DOMAINE disques optiques
L1 TERM mastering	L2 TERME gravure
DEFINITION The process of encoding digital data on a glass master disc prior to CD replication.	DÉFINITION Une étape dans la réalisation d'un CD-ROM qui consiste à enregistrer des données sur un disque maître en verre et qui précède le processus de duplication.
EXAMPLE Mastering produces the glass master disc that is the first generation in a multi-step process that produces the stamplers that actually press the digital pattern into a disc.	EXEMPLE Gravure : il s'agit du pressage, par enregistrement optique, des données issues de l'étape précédente, sur un disque ''maître'' en verre, recouvert de la couche sensible.

FIGURE 1 : Représentation bilingue d'une notion dans une banque de terminologie traditionnelle (notion = *mastering/gravure*).

Une telle fiche ne satisfera pas tous les usagers. En effet, nombreux sont ceux qui préféreraient sans doute obtenir une perspective élargie de deux catégories de renseignements : la définition et le contexte. La définition sert à éclaircir le sens d'un terme en expliquant comment la notion correspondante s'insère dans les réseaux notionnels du domaine d'étude. Pour comprendre une notion spécialisée, il faut comprendre les rapports qui l'associent aux autres notions du même domaine. Malheureusement, une

3. Cet exemple s'inspire de la structure de TERMIUM III, la banque de terminologie de Services gouvernementaux Canada (quoique nous n'utilisons pas la véritable fiche TERMIUM pour *gravure* à titre d'exemple). Évidemment, d'autres modèles que celui-ci existent déjà ou sont en cours d'élaboration. Mais ils sont trop nombreux pour être décrits ici. Pour un exposé de la problématique associée à la conception des banques de terminologie (y compris le volet notionnel), voir Sager 1990.

simple définition (même si elle est bien construite, ce qui n'est pas toujours le cas[4]) ne donne souvent qu'un aperçu très sommaire – une « perspective » réduite – des réseaux notionnels du domaine en question. Même si la définition laisse entrevoir les notions connexes, il n'en reste pas moins que la notion définie demeure fondamentalement isolée. En effet, l'usager n'a aucun moyen de naviguer dans le réseau notionnel (la « salle de bal ») qui pourrait s'y rattacher. Dans la première partie de la présente communication, nous montrerons comment le modèle à base de connaissances proposé par les chercheurs en intelligence artificielle peut offrir une perspective élargie des structures notionnelles.

Les contextes, deuxième catégorie de renseignements figurant souvent sur les fiches terminologiques, peuvent aussi produire une perspective réduite. Des deux grandes fonctions suivantes, ils peuvent remplir soit l'une, soit l'autre, soit un agencement des deux : d'une part compléter la définition par l'ajout de renseignements notionnels[5], d'autre part, illustrer l'usage d'un terme (caractéristiques grammaticales, cooccurrents etc.). Or, les contextes que l'on trouve dans les banques de terminologie n'offrent eux aussi qu'une perspective réduite du sens et de l'usage des termes, et le manque d'espace en est partiellement responsable. En outre, les banques de terminologie traditionnelles n'illustrent qu'une fraction des tournures auxquelles se prêtent les termes en contexte. Dans la seconde partie de l'article, nous montrerons qu'il est possible d'obtenir une perspective élargie de l'information en associant une banque de terminologie à un vaste corpus de textes spécialisés, accessible par l'intermédiaire d'un analyseur de corpus[6].

Perspectives élargies des structures notionnelles : une démarche fondée sur les connaissances

De nombreux chercheurs ont souligné l'importance d'élargir le contenu notionnel des sources traditionnelles de renseignements terminologiques (Ahmad *et al.* 1989 ; Blampain *et al.* 1992 ; Kukulska-Hulme et Knowles 1989 ; Pearson *et al.* 1993, etc.). Nos propres recherches[7] veulent montrer que les technologies et les techniques de pointe associées au génie cognitif (sous-domaine de l'intelligence artificielle axé sur la modélisation informatique des domaines de la connaissance) permettent d'envisager la prochaine génération de banques de terminologie comme une sorte d'hybride entre la banque traditionnelle et la base de connaissances, comme on l'appelle en intelligence artificielle. Ce modèle, nous l'appelons *base de connaissances terminologiques (BCT)*. Depuis 1989, nous effectuons nos recherches à l'aide d'un outil appelé *CODE* (*Conceptually Oriented Description Environment*), mis au point à l'Université d'Ottawa, et basé

4. Eck 1993, et Eck et Meyer 1993 analysent certains problèmes couramment associés aux définitions suggérées par les banques de terminologie, notamment le manque d'homogénéité (entre les notions coordonnées, c'est-à-dire celles qui partagent la même notion générique) des termes génériques et des traits notionnels. Ces problèmes empirent lorsque les définitions sont extraites d'un contexte et non pas rédigées par les terminologues eux-mêmes (par exemple, les définitions de notions coordonnées peuvent provenir de sources variées et, par le fait même, d'auteurs différents).
5. En effet, faute de pouvoir trouver un contexte définitoire, les banques de terminologie remplacent parfois la définition par un contexte général. Dans TERMIUM III, ces contextes se rangent dans deux catégories : CONT indique que le contexte est suffisamment détaillé pour cerner la notion, alors que EX annonce que le contexte ne l'est pas assez.
6. L'analyse de corpus est un secteur de la technologie linguistique en plein essor. Le manque d'espace nous empêche toutefois de décrire les différents types de logiciels en cours d'élaboration. Les thèmes que nous abordons ici s'inspirent en grande partie de l'expérience que nous avons acquise avec *TACT*, un concordancier mis au point par le *Centre for Computing in the Humanities* de l'Université de Toronto (Canada). On peut se le procurer facilement via Internet. Sinclair 1991 et Svartvik 1992 donnent un aperçu de l'utilisation des concordanciers et des analyseurs de corpus en recherche linguistique.
7. Bowker et Meyer 1993, Eck et Meyer 1993, Meyer 1992, Meyer *et al.* 1992 et 1994, Miller, Meyer et Michaud 1991, etc.

sur des recherches en génie cognitif. Cet outil nous a servi à construire un prototype de BCT dans le domaine du stockage optique : le système COGNITERM.

Il serait superflu de donner ici une description détaillée de notre technologie et de notre méthodologie. Pour l'essentiel, notre démarche consiste à encoder les principaux caractères de chaque notion, qu'il s'agisse d'attributs notionnels (propres à la notion elle-même), de relations (avec d'autres notions) ou de fonctions. Pour ce faire, nous avons conçu des « types sémantiques » suivant un modèle notionnel et d'après les renseignements trouvés dans la documentation écrite ou fournis par des experts. L'exemple de la figure 2, tiré de notre BCT, montre que le type sémantique *processus de production* se caractérise par des relations hiérarchiques (générique-spécifique et méronymiques[8] [partie-ensemble]), des relations non hiérarchiques et une fonction.

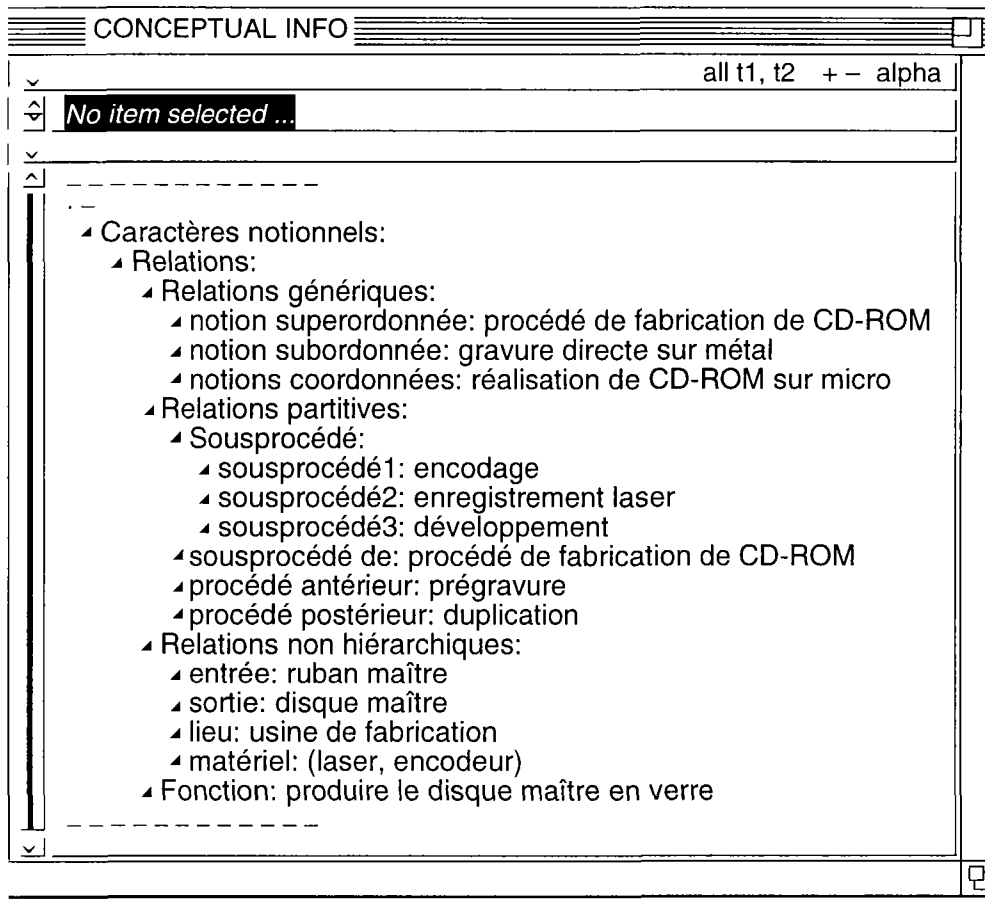


FIGURE 2 : Représentation des caractères notionnels de *gravure* dans la base de connaissances.

8. Chaffin *et al.* ont classé les relations méronymiques en trois grandes catégories (qui comptent aussi de nombreuses sous-catégories) : partie-ensemble (tasse-anse), matériaux (tasse-porcelaine) et phases (croissance-adolescence). Les relations qui caractérisent nos « sous-procédés » par rapport au « procédé de production » correspondent à la catégorie « phase » de Chaffin *et al.*

L'essence même des renseignements fournis par cette description notionnelle n'a rien pour ébahir les terminologues. En effet, ils auraient sans doute acquis les mêmes connaissances en préparant une fiche pour une banque de terminologie traditionnelle. D'ailleurs, on imagine mal comment un terminologue pourrait choisir ou rédiger une définition pour *gravure* sans savoir qu'il s'agit d'une étape du procédé de fabrication de CD-ROM (notion superordonnée), qu'un ruban maître (entrée) sert à produire le disque maître (sortie), que la gravure compte trois sous-procédés, à savoir l'encodage, l'enregistrement laser et le développement, etc. Or, les renseignements notionnels de notre modèle de fiche se caractérisent non pas par leur essence même, mais par deux autres aspects : 1) ils apparaissent bel et bien sur la fiche terminologique « officielle » (habituellement, les terminologues emmagasinent plutôt ces renseignements dans leur mémoire ou prennent quelques notes pour leur usage personnel seulement) ; 2) ils sont présentés suivant une structure explicite et facile à comprendre. Cette structure s'associe aux nombreuses autres fonctions du système *CODE,* qui assurent l'homogénéité de l'information dans la base de connaissances. Résultat : l'usager peut obtenir une perspective intéressante de l'information, largement supérieure à celle qu'offrent les définitions et les contextes proposés par les banques de terminologie traditionnelles.

En outre, *CODE* permet à l'usager d'obtenir une perspective très importante de l'information : la « vue d'ensemble » d'un ou de plusieurs réseaux notionnels. La figure 3, par exemple, offre une vue d'ensemble d'un champ notionnel de notre BCT pour le domaine du stockage optique, en l'occurrence les relations hiérarchiques qui caractérisent l'« environnement notionnel » de *gravure.* Comme l'illustre la figure 3, il s'agit de relations partitives que nous appelons ici « sous-procédés ». Ainsi, l'usager désirant en savoir plus sur la gravure apprendra qu'il s'agit d'une étape du *processus traditionnel de réalisation de CD-ROM,* et que la gravure comporte elle-même trois sous-étapes. Si l'usager voulait aller encore plus loin, il apprendrait que le *processus traditionnel de réalisation de CD-ROM* se distingue de la *réalisation de CD-ROM sur micro-ordinateur.* Et si cette dernière notion lui était inconnue, il pourrait l'éclaircir à l'aide de la base de connaissances et obtenir une structure comparable à celle qu'illustre la figure 2.

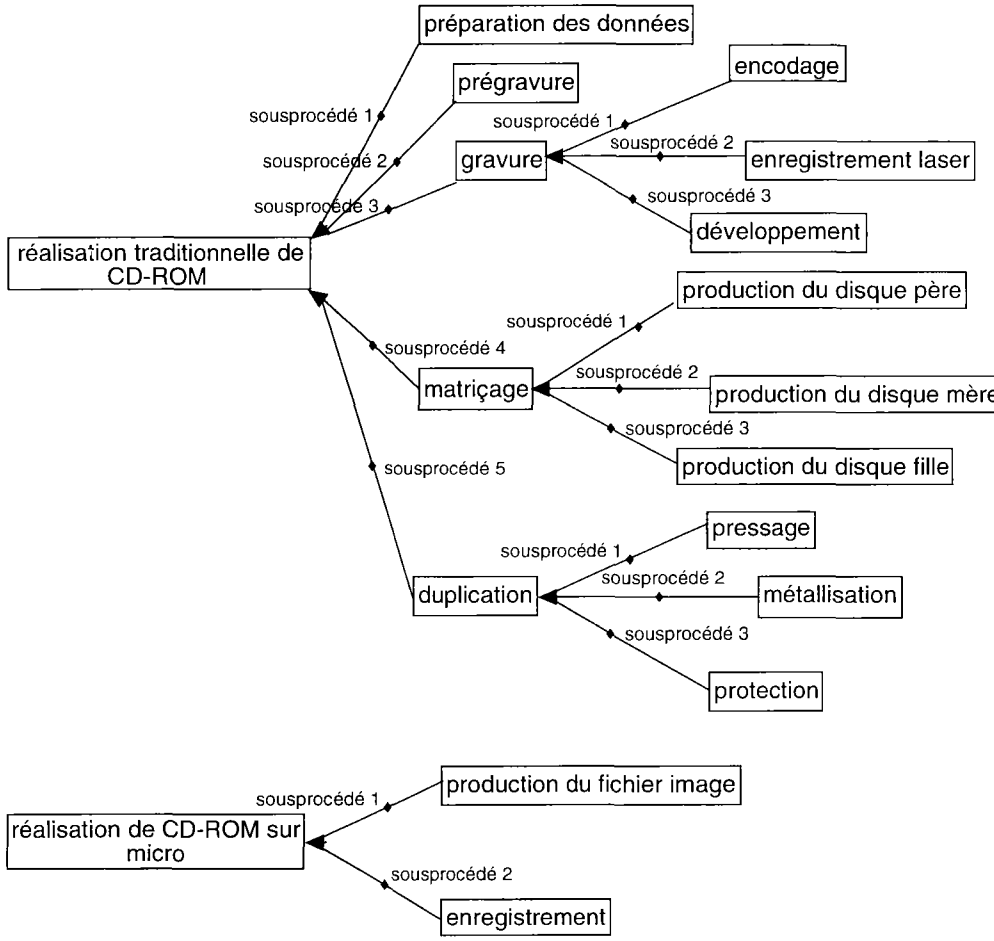


FIGURE 3 : Relations partitives (sous-procédés) associées à la notion *gravure*.

Les relations hiérarchiques, peu importe leur importance ou leur prédominance dans le domaine du stockage optique, ne sont pas les seules relations permettant d'éclaircir le sens d'une notion. Comme le montre la figure 4, le système *CODE* produira automatiquement (à partir des renseignements donnés par la description notionnelle illustrée à la figure 2) le graphe des relations non hiérarchiques auxquelles s'associe la notion *gravure*. La figure 4 montre que l'on effectue la gravure à un lieu (nom de la relation) précis que l'on appelle *usine de fabrication* (valeur de la relation), que les données sont entrées à l'aide d'un *ruban maître,* que le résultat donne un *disque maître* (sortie), et ainsi de suite. Lorsque les valeurs des relations sont elles-mêmes des notions spécialisées (c'est le cas de toutes les valeurs illustrées à la figure 4), on peut obtenir pour chacune une description notionnelle comparable à celle qu'illustre la figure 2 et, de là, demander au système de produire des graphes comme

ceux des figures 3 et 4. Ainsi, l'usager peut danser à loisir dans une salle de bal sans cesse grandissante de relations notionnelles présentées sous forme de texte (figure 2) ou de graphe (figures 3 et 4).

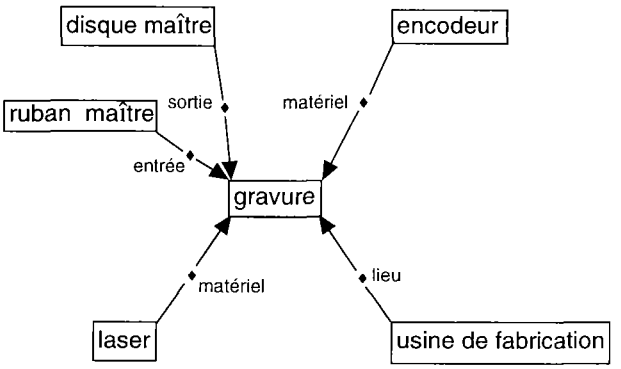


FIGURE 4 : Relations non hiérarchiques comprises dans l'environnement notionnel de *gravure*.

En conclusion, la base de connaissances offre une perspective beaucoup plus large et cohérente des structures notionnelles que ne le font les définitions proposées par les banques de terminologie traditionnelles. Ces définitions, comme l'ont souligné Kukulska-Hulme et Knowles, ne donnent qu'un aperçu fragmentaire du domaine, un peu comme « un puzzle qu'on n'arrive pas à assembler parce qu'il y a des pièces qui manquent, et qu'on n'a pas l'image de l'ensemble qui devrait servir de guide » (Kukulska-Hulme et Knowles 1989 : 382). Mais ce n'est pas tout ; nous avons avancé (Eck 1993 ; Meyer et Eck 1993) que le modèle à base de connaissances pouvait aider les terminologues à rédiger d'emblée de meilleures définitions. À cette fin, nous avons doté le système *CODE* d'une fonction comparative capable de produire (à partir de descriptions notionnelles comme celles de la figure 2) un tableau des caractères distinctifs ressemblant à une feuille de calcul (voir figure 5). Ce tableau aidera non seulement le terminologue à rédiger ses définitions, mais il lui donnera aussi, une fois de plus, une perspective plus large des structures notionnelles que ne le font les banques de terminologie traditionnelles.

Characteristic Comparison Matrix

	▷ procédé de fabrication de CD-ROM: (s,s)	
	☐ réalisation traditionnelle de CD-ROM	☐ réalisation de CD-ROM sur micro
☐ entrée	données numériques	(disque CD-R, données numériques
☐ sortie	CD-ROM	disque lisible par un lecteur CD-ROM
☐ lieu	dans une usine de fabrication	sur place (à l'aide d'un micro-ordinateur)
☐ matériel	le matériel utilisé dans les sousprocédés	(lecteur CD-R, disque CD-R, micro-ordinateur)
☐ Fonction	produire un nombre élevé de disques CD-ROM	enregistrer un disque CD-R, qui tournera sur un lecteur CD-ROM
☐ sousprocédé	préparation des données	production du fichier image
☐ sousprocédé	prégravure	enregistrement
☐ sousprocédé	gravure	
☐ sousprocédé	matriçage	
☐ sousprocédé	duplication	

FIGURE 5 : Ce tableau comparatif illustre les caractères distinctifs des notions *réalisation traditionnelle de CD-ROM* et *réalisation de CD-ROM sur micro.* Les rangs donnent les traits distinctifs, et les colonnes, les notions comparées. Les valeurs des caractères apparaissent dans les cellules.

Perspectives élargies de contextes terminologiques : l'utilisation de corpus

Comme nous l'avons avancé dans la première partie de la présente communication, une autre conséquence de l'« effet de focalisation », perspective réduite offerte par les banques de terminologie traditionnelles, est le potentiel limité des contextes d'utilisation. Bien que ces contextes servent à illustrer d'une part le sens et d'autre part l'usage, nous insisterons particulièrement sur l'usage. Si nous affirmons que le potentiel des contextes d'utilisation est « limité », c'est qu'il ne permet pas d'illustrer pleinement l'usage et ce, pour deux raisons. D'abord parce que tout contexte est limité, car il n'est en fait qu'*un fragment d'un texte plus long, duquel il a été extrait artificiellement.* On ne peut jamais avoir la certitude que cette extraction artificielle ne nuira pas à l'usager ou à l'exécution d'une tâche. Voici ce que disait Sinclair (1991 : 5) à cet égard :

> Les mots sont tous en relation avec leur contexte. Dans chaque fragment d'un texte, le choix des mots s'inspire, dans certains cas, des choix que l'auteur a faits ailleurs dans le texte. C'est pourquoi un exemple n'est jamais complet s'il ne s'agit pas d'un texte entier. (Traduction libre[9])

En outre, si les contextes offerts par les banques de terminologie ont un potentiel fondamentalement limité, c'est aussi parce que *les terminologues tâchent de choisir des contextes qui, selon eux, satisferont le plus grand nombre d'usagers.* Évidemment, les terminologues ne peuvent avoir qu'une vague idée des personnes qui utiliseront la terminologie et de l'usage qu'elles en feront. Les usagers se distinguent en effet par leur degré de familiarité avec le domaine (depuis les novices jusqu'aux experts établis) et par l'usage qu'ils feront de l'information (traduction, rédaction ou simple apprentissage[10]).

De nos jours, les lexicographes ont accès à des masses croissantes de données textuelles, et les analyseurs de corpus évoluent rapidement. Cela ouvre des perspectives tout à fait nouvelles pour la recherche terminologique : l'étude de moyens permettant d'associer la description d'un terme à un corpus électronique. Comme le souligne Kukulska-Hulme 1990, « [...] le fossé qui sépare dictionnaires et textes n'a jamais été si étroit ». D'une part, les terminologues sont parvenus à rapprocher dictionnaires et textes avant les lexicographes, car ils s'inspirent depuis toujours d'exemples réels alors que les lexicographes ont pris l'habitude de se fier en grande partie à leur instinct, pratique qui a récemment soulevé un débat mouvementé dans le milieu de la lexicographie[11]. D'autre part, les lexicographes sont en avance sur les terminologues dans le domaine de l'analyse de corpus électroniques (d'où le terme anglais bien connu *corpus lexicography*). En effet, dans la pratique, les terminologues ont besoin de consulter une vaste gamme de textes spécialisés, même pour de petites recherches thématiques. Or, trouver ces textes sur support informatique n'est pas encore chose facile, surtout lorsqu'il s'agit de textes très récents qui doivent servir à définir des néologismes. Évidemment, les choses évoluent rapidement, et l'heure est venue, selon nous, d'aborder l'analyse de corpus dans une perspective terminologique.

L'analyse de corpus axée sur la terminologie est un domaine dans lequel notre expérience demeure, pour le moment, très limitée. La problématique que soulèvera ce nouveau domaine de recherche ne peut donc faire l'objet d'un traitement approfondi dans la présente communication[12]. Nous nous contenterons seulement d'illustrer globalement certaines « perspectives » (pour garder l'image du judas et de la salle de

---

9. Voici la citation originale : « *Any instance of language depends on its surrounding context. The details of choice shown in any segment of a text depend – some of them – on choices made elsewhere in the text, and so no example is ever complete unless it is a whole text.* »

10. On n'a qu'à penser à l'étudiant en médecine qui chercherait des termes spécialisés à la lecture d'un manuel de cours. Nous pensons en effet que le potentiel pédagogique des banques de terminologie n'a pas reçu l'attention qu'il mérite en recherche terminologique. Selon nous, ce potentiel inexploité s'avère un excellent motif pour améliorer la représentation des structures notionnelles dans les sources de renseignements terminologiques.

11. La publication du dictionnaire *Cobuild* (Sinclair *et al.* 1988), fruit du premier travail lexicographique associé de très près à un vaste corpus électronique, est le principal catalyseur de ce débat. En effet, la lexicographie a toujours donné préséance à l'instinct des lexicographes, d'où l'utilisation d'exemples créés de toutes pièces dans les dictionnaires de langue générale. Sinclair (1991) brosse un tableau plus détaillé des différentes sources de renseignements lexicographiques.

12. En bref, nous croyons que l'analyse de corpus axée sur la terminologie, par rapport à l'analyse de corpus axée sur la lexicographie, soulèvera une problématique quelque peu différente, surtout en raison de la masse de connaissances qu'elle apportera à la recherche terminologique.

bal) qu'offrent les corpus pour montrer l'utilisation des termes en contexte. Pour analyser des corpus, les lexicographes et les terminologues auront tout avantage à utiliser un concordancier. Les figures 6 et 7 montrent la présentation type des index de mots-clés en contexte produits par ces logiciels. Évidemment, ces exemples n'illustrent qu'une fraction des entrées que nous avons obtenues, car notre corpus s'élève à un million de mots pour le seul domaine du stockage optique. L'échantillon reproduit à la figure 6 montre d'intéressants contextes phraséologiques antéposés pour les termes *master* et *mastering*. On peut y déceler des cooccurrents de *master (make, create, cut, prepare, write onto)* et des composés (*gold master, CD master, glass master, desktop mastering, do-it-yourself mastering*). Quant à l'échantillon reproduit à la figure 7, il montre d'intéressants contextes phraséologiques postposés, notamment des composés (*master floppies, master mold, master nickel cookie cutter, mastering facility, mastering house, mastering unit*) et certains usages grammaticaux (*master on a computer, master to CD-ROM*).

to make a CD-ROM	master	and run off copies
100 lets users create	master	CD-ROM disks in-house to
In contrast, cutting a	master	CD-ROM through a service
$2,000 to prepare the	master	Consts compare very
writes a file onto the	master	Typing LLINSTALL at the
data is written onto	master	disks using special record
pressing duplicates of a	master	disc (much like records) at
bureau can press a gold	master	that users check for errors
and desktop	mastering	may be the key to
authoring with a desktop	mastering	system prices at less
the price of desktop	mastering	systems to come down
a " do-it-yourself "	mastering	and duplication system
able to buy a CD-ROM	mastering	device for less than
the expensive outside	mastering	process altogether
Free first tape	mastering	plus 50 replicas with
just provides a CD	master	the hardware company does
create a traditional glass	master	or stamped disc for each
to produce a trial	master	before the replication
to produce encrypted CD-ROM	master	data. The information can

FIGURE 6 : Index de mots-clés en contexte pour *master* et *mastering* : contextes antéposés intéressants.

$250 to $500 to cut a	master	compact disk, and orders
it and preparing the	master	disc or any other storage
However, the 1980's two	master	Disk Cache Processors
from a mold created from a	master	disk. The original master
Before you begin, copy the	master	floppies and use a copy
process, a two-part	master	mold is made from an
or creation of the	master	nickel "cookie cutter," is
a contractor who produces a	master	tape and stamps out the
employees to put together	master	information for the disks
it hired a contractor to	master	its regulations for CD-ROM
and then sent to a	mastering	facility for replication
to an outside	mastering	factory, where a CD
and to send the image to a	mastering	plant for volume
having to use an outside	mastering	service, and it also
provides. While most	mastering	houses do offer
$30,000 made sense when	mastering	hardware cost upward
Inc.'s CD-ROM	mastering	unit (Hardware Review)
part of it away. CD-ROM	mastering	machines and WORM drives
improvements in the	mastering	process.
Full Text: CD	mastered	on NEXT computer
companies are finding that	mastering	data to CD-ROM is

FIGURE 7 : Index de mots-clés en contexte pour *master* et *mastering* : contextes postposés intéressants.

Comme nous l'avons vu, les concordanciers sont capables de détecter une *quantité* énorme de contextes phraséologiques. Pour *master* (verbe et substantif), nous avons obtenu un index d'environ 300 entrées à partir d'un corpus anglais d'un million de mots. De plus, l'usager peut, au besoin, agrandir le contexte d'un terme jusqu'à concurrence du texte au complet. La figure 8, par exemple, illustre une situation où l'usager s'interrogeait sur l'expression « *write to master disks* » apparaissant à la deuxième ligne de l'index. Il a donc demandé un contexte élargi pour *master*, en l'occurrence trois lignes avant et trois lignes après le terme.

KWIC VIEW

writes a file onto the	master	Typing LLINSTALL at the
data is written onto	**master**	disks using special
pressing duplicates of a	master	disc (much like records) at

EXPANDED VIEW

CD-ROM, which stands for compact disk read-only memory, is fundamentally an adaptation of the compact disc digital audio recording system. As with audio CDs, digital data is written onto **master** disks using special recording equipment that makes microscopic pits in the disk surface. The information encoded in the pits can be read simply by detecting changes in reflectivity: The pits are darker than the shiny silver background of the disk.

FIGURE 8 : Élargissement d'un contexte.

Ce court exemple, associé au contexte élargi reproduit à la figure 8, aidera sans doute l'usager à mieux cerner les contextes dans lesquels l'expression « *write to a master disk* » s'insère naturellement. Il verra, entre autres, qu'un sujet typique serait « *digital data* ». Bref, cet exemple et ceux des figures 6, 7 et 8 illustrent différents usages caractérisant les termes *master* (verbe et substantif) et *mastering*. De toute évidence, ils illustrent l'usage du terme suivant une « perspective » beaucoup plus élargie que ne le font les banques de terminologie traditionnelles. Cependant, comme nous l'avons déjà mentionné, illustrer l'usage n'est qu'une fonction des contextes proposés par ces banques de terminologie ; l'autre, c'est d'illustrer le sens. Le contexte élargi de la figure 8, par exemple, fournirait à l'usager des indices importants sur le sens de l'expression *write to a master disk,* et un contexte encore plus élargi lui fournirait de plus amples précisions. Mais cette constatation outrepasse l'objet de la présente partie, qui porte sur les contextes d'utilisation comme moyen d'illustrer l'usage. Elle nous amène à aborder l'analyse notionnelle, ce qui ferme la boucle et nous reporte à la partie précédente sur la représentation des notions dans les bases de connaissances. Notre conclusion traitera brièvement de certains rapports que nous établissons entre les bases de connaissances et les corpus spécialisés. Autrement dit, nous expliquerons comment les deux « perspectives » que nous avons abordées peuvent s'harmoniser pour offrir une perspective intégrée des notions et des termes présents dans notre « salle de bal » terminologique.

Harmonisation des modèles à base de connaissances et à base de corpus

Notre modèle de représentation des données terminologiques englobe deux éléments qui, à première vue, semblent plutôt incompatibles : une base de connaissances, d'une part, qui se caractérise par l'*explicitation* très évidente de ses données terminologiques, et un vaste corpus, d'autre part, qui se distingue par l'*implicitation* de son contenu. Notre démarche veut montrer que, même si les bases de connaissances et les corpus se situent à des extrémités opposées de l'axe explicitation-implicitation, ils n'en sont pas moins deux modèles complémentaires de représentation des données terminologiques. Dans les prochaines parties de notre exposé, nous examinerons brièvement cette complémentarité sous deux angles : les types de notions et les types d'usagers et de tâches.

Types de notions

Les terminologues savent très bien que les notions spécialisées ne se situent pas toutes au même niveau. D'ailleurs, le degré de « maturité » des notions est un facteur important de variation. Comme l'expliquent Ahmad *et al.* (1992 : 147) :

> [...] la connaissance du domaine comporte un cycle de vie bien à elle : naissance (avènement d'un domaine nouveau ou changements dans un domaine existant), maturation (croissance et évolution), maturité (adoption de méthodes de travail et diffusion des connaissances par des livres, des revues, des travaux, des articles, des documents commerciaux, etc.), mutation (à même le domaine ou par contact avec d'autres domaines) et, finalement, mort (obsolescence). La terminologie évolue de la même façon au sein de divers milieux de communication, et ce n'est qu'au moment où un terme devient « mature » que la normalisation devient fondamentale, voire possible. (Traduction libre)

Les banques de terminologie traditionnelles tiennent compte, dans une certaine mesure, du degré de maturité des notions. Par exemple, l'usager de TERMIUM III remarquera probablement que les notions matures sont « étayées textuellement » par des contextes définitoires alors que les notions moins évoluées sont éclairées par différents types de contextes non définitoires[13]. Parallèlement, dans notre modèle, les notions les plus matures sont celles dont l'« environnement » notionnel se prête le mieux à la structure d'une base de connaissances. De fait, la représentation des données dans une base de connaissances à structure très formalisée se révélera utile pour les notions ayant fait l'objet d'un consensus assez global chez les experts du domaine ou dans les cas où la normalisation s'impose pour des raisons pratiques[14]. Quant aux notions tout à fait nouvelles, nous pensons que la meilleure solution n'est pas de les consigner dans une base de connaissances, mais de laisser l'usager naviguer à sa guise dans un corpus pour qu'il puisse voir les diverses façons dont elles y sont traitées au premier abord. Et pour les notions comprises entre ces deux extrêmes, la combinaison des deux modèles s'avérerait peut-être la meilleure solution.

Types d'usagers et de tâches

Comme nous l'avons signalé, lorsque les terminologues préparent une fiche pour une banque de terminologie traditionnelle, ils émettent inévitablement un certain nombre d'hypothèses sur les besoins de l'usager « moyen ». Évidemment, cette méthode offre des perspectives restreintes, car les usagers diffèrent à de nombreux égards, notamment en ce qui concerne la connaissance du domaine et la tâche qui donne lieu à la recherche terminologique.

Connaissance du domaine. Selon nous, aucune étude n'a cherché à déterminer avec exactitude ce que les différents types d'usagers veulent savoir lorsqu'ils interrogent une banque de terminologie. Néanmoins, il se peut fort bien que moins l'usager s'y connaît dans un domaine, plus il comptera sur la banque de terminologie pour l'éclairer. On n'a qu'à penser aux traducteurs (plus grands consommateurs de terminologie informatisée au Canada) qui, avant d'entreprendre la traduction d'un texte, doivent avoir acquis une certaine compréhension du domaine dans lequel ils auront à travailler. Pour ce faire, ils devront sans doute passer beaucoup de temps à consulter une foule de documents, car la connaissance d'un domaine (comme l'a montré la section sur les perspectives élargies des structures notionnelles) ne s'acquiert pas facilement à l'aide d'une banque de terminologie traditionnelle. Dans la deuxième partie, nous avons prétendu que le modèle à base de connaissances permettrait à l'usager d'acquérir des connaissances dans son domaine de recherche. Or, les terminologues chargés d'alimenter la base de connaissances ne peuvent songer à tous les renseignements dont aura besoin l'usager pour comprendre ou rédiger un texte convenablement ; la base de connaissances n'offrira en effet qu'une certaine facette du domaine. À notre avis, le corpus deviendrait alors un prolongement de la base de connaissances ; il permettrait à l'usager en quête de plus amples renseignements d'effectuer ses recherches à sa guise.

13. Dans TERMIUM III, par exemple, les contextes non définitoires se rangent dans deux catégories : CONT (contexte) et EX (exemple d'utilisation).

14. Par exemple, les notions et les termes associés à la fabrication d'un produit en usine doivent être normalisés à l'amorce du processus d'élaboration pour que la mise en fabrication du produit (y compris la rédaction des documents nécessaires) se déroule sans encombre.

Tâches. L'usager se servira de données terminologiques pour comprendre, rédiger ou traduire un texte. Comme nous l'avons mentionné, la base de connaissances aidera l'usager à comprendre un texte. Le corpus l'aidera aussi dans cette démarche, mais il lui sera encore plus utile pour rédiger. L'élaboration de techniques d'utilisation de corpus adaptées aux besoins des usagers offre d'ailleurs de vastes avenues de recherche, car il ne s'agit pas de déterminer « comment s'énoncent les concepts dans un domaine », mais plutôt « comment s'énoncent les concepts dans certains *types de textes* et dans certaines *situations de communication* propres à un domaine ». Par exemple, l'expert s'adressera probablement à un pair en des termes fort différents de ceux qu'il utiliserait avec un novice. Et d'un extrême à l'autre, les situations de communication sont encore plus nombreuses.

En conclusion, la base de connaissances et le corpus se situent aux antipodes de l'axe explicitation-implicitation auquel correspond un axe analogue pour les types de notions et pour les types d'usagers et de tâches. Or, à l'une ou l'autre extrémité de cet axe, les banques de terminologie traditionnelles ne peuvent s'avérer suffisamment utiles. D'une part, elles contiennent des données terminologiques (et surtout notionnelles) trop implicites pour convenir parfaitement au traitement des notions très matures ou aux usagers en quête de renseignements sur un domaine, surtout de renseignements nécessaires pour comprendre un texte. D'autre part, les contextes qu'elles proposent pour illustrer l'usage d'un terme sont trop rigides (parce qu'ils ont été extraits artificiellement d'un texte complet) pour convenir parfaitement aux notions nouvelles ou aux usagers en quête de renseignements qui les aideront à rédiger.

Conclusion

En consultant une banque de terminologie traditionnelle, l'usager veut obtenir des « perspectives » à la fois linguistiques et notionnelles de l'information. Or, les banques de terminologie, du fait de leur conception même, ne leur offrent en réalité que des « perspectives réduites ». D'une part, les définitions qu'elles proposent ne révèlent qu'une partie des structures notionnelles, ce qui limite leur utilité à la normalisation de notions matures et à la transmission de renseignements sur un domaine. D'autre part, les contextes d'utilisation qu'on y trouve sont extraits artificiellement de véritables textes et, comme ils sont peu nombreux, ils ne peuvent illustrer tous les usages possibles d'un terme en fonction de toutes les situations de communication propres à un domaine. Pour amoindrir l'« effet de focalisation » imposé par ces contraintes, nous proposons un modèle de représentation des données terminologiques qui, en combinant une base de connaissances et un corpus, permet à l'usager d'obtenir des « perspectives élargies » de l'information. À notre avis, c'est un modèle dynamique : la base de connaissances conviendra davantage pour les notions matures et dans les cas où l'usager voudra obtenir une foule de renseignements dans un domaine particulier ; le corpus conviendra pour les notions récentes et dans les cas où l'usager devra exécuter des tâches axées surtout sur la rédaction.

Les technologies nécessaires à la réalisation du modèle que nous proposons pour la conception des banques de terminologie sont déjà bien au point et évoluent rapidement. Pour nos recherches, nous avons construit un prototype de base de connaissances terminologiques avec les matériaux qu'offre la technologie du génie cognitif. Or, même si les analyseurs de corpus sont facilement accessibles, les terminologues

ont pris du retard sur les lexicographes, notamment parce qu'il s'est avéré plus difficile de trouver des corpus de textes récents et spécialisés que des corpus de textes généraux. Mais ce n'est qu'une question de temps. Si les perspectives de représentation des données terminologiques s'élargissent suivant notre modèle, la disponibilité de la technologie n'en sera pas l'unique raison. En effet, il faut accorder autant de crédit au fait que nos orientations s'harmonisent avec les méthodes que les terminologues utilisent depuis toujours : même avant l'avènement des outils de développement, ils voyaient l'analyse notionnelle comme un excellent moyen d'acquérir des connaissances dans un domaine. Et bien avant l'apparition des corpus électroniques et des analyseurs de corpus, la documentation spécialisée demeurait leur principale source de renseignements terminologiques.

Remerciements

Nous aimerions remercier René Morin pour avoir traduit et adapté la version anglaise de la présente communication ainsi que Jean Quirion pour avoir commenté la version française. Nous remercions également Lynne Bowker, Karen Eck, Kristen Mackintosh et Douglas Skuce pour leur importante contribution. Le projet COGNITERM est financé par le Conseil de recherches en sciences humaines du Canada et par le Service de la recherche de l'Université d'Ottawa. L'élaboration du système *CODE* a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada, Bell-Northern Research et le Fonds d'encouragement à la recherche dans les universités du gouvernement de l'Ontario.

Références

AHMAD, K., FULFORD, H. et M. ROGERS (1992) : « The Elaboration of Special Language Terms: the Role of Contextual Examples, Representative Samples and Normative Requirements », *Proceedings of the Fifth EURALEX International Congress*, Part I, Tampere, pp. 139-149.

BLAMPAIN, D., PETRUSSA, P. et M. VAN CAMPENHOUDT (1992) : « À la recherche d'écosystèmes terminologiques », A. Clas et H. Safar (dir), *L'environnement traductionnel: La station de travail du traducteur de l'an 2001*, Sillery, Presses de l'Université du Québec et AUPELF-UREF, actualité scientifique, pp. 273-282.

BOWKER, L. et I. MEYER (1993) : « Beyond Textbook Concept Systems: Handling Multidimensionality in a New Generation of Term Banks », *Proceedings of the Third International Congress on Terminology and Knowledge Engineering (TKE '93)*, pp. 123-137.

CHAFFIN, R., HERRMANN, D. et M. WINSTON (1988) : « An Empirical Taxonomy of Part-Whole Relations: Effect of Part-Whole Relation Type on Relation Identification », *Language and Cognitive Processes*, 3, 1, pp. 17-48.

ECK, K. et I. MEYER (1993) : « Definition Construction in a Terminological Knowledge Base », *ASTM Symposium on Standardizing and Harmonizing Terminology: Theory and Practice*, Philadelphia, October 1993.

ECK, K. (1993) : « Bringing Aristotle into the 20th Century: Definition-Oriented Concept Analysis in a Terminological Knowledge Base », mémoire, École de traduction et d'interprétation, Université d'Ottawa, Ottawa, Canada.

HUMBLEY, John (à paraître) : « Exploitation d'un vocabulaire combinatoire : syntaxe, phraséologie, analyse conceptuelle », *Terminologies nouvelles.*

KUKULSKA-HULME, A. (1990) : « A Dictionary View of Technical Writing », *Computers and Writing III,* 5-7 avril 1990, Heriot-Watt University, Edinburgh.

KUKULSKA-HULME, A. et F. KNOWLES (1989) : « L'organisation conceptuelle des dictionnaires automatiques pour textes techniques », *META,* 34-3, pp. 381-397.

MEYER, I. (1992) : « Knowledge Management for Terminology-Intensive Applications: Needs and Tools », J. Pustejovsky et S. Bergler (dir), *Lexical Semantics and Knowledge Representation,* Berlin, Springer Verlag, pp. 21-37.

MEYER, I., BOWKER, L. et K. ECK (1992) : « COGNITERM: an Experiment in Building a Knowledge-Based Term Bank », *Proceedings of the Fifth EURALEX International Congress,* Tampere, pp. 159-172.

MEYER, I., ECK, K. et D. SKUCE (à paraître) : « Systematic Concept Analysis Within a Knowledge-Based Approach to Terminology », *A Handbook of Terminology,* S. E. Wright et G. Budin (dir), Amsterdam/Philadelphia, John Benjamins.

MILLER, D., MEYER, I. et D. MICHAUD (1991) : « Terminologie et analyse notionnelle assistée par ordinateur », *Actes du Colloque international sur les industries de la langue, Tome II,* Québec, Office de la langue française et STQ, pp. 781-800.

MOSER-MERCER, Barbara (1987) : « Man/Machine Interface in Translation and Terminology », *META,* 32-2, pp. 156-163.

PEARSON, J. *et al.* (1993) : *Terminology and Extra-Linguistic Knowledge,* Report 2 of the ET10/66 Consortium (Dublin City University, CRP-CU - Luxembourg, ALTEC -Lisbon, INCOM - Bonn).

SAGER, Juan (1990) : *A Practical Course in Terminology Processing,* Amsterdam/Philadelphia, John Benjamins.

SINCLAIR, J. (1991) : *Corpus, Concordance, Collocation,* Oxford University Press.

SINCLAIR, J. *et al.* (1987) : *Collins COBUILD English Language Dictionary,* London, Collins.

SVARTVIK, Jan (dir) (1992) : *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82,* Stockholm 4-8 août 1991, Berlin, Mouton de Gruyter.

29

Les aspects terminologiques de la traduction : évolution des outils logiciels

Élisabeth BLANCHON

CTN, CNRS, Paris, France

• *Abstract* •

Terminology is a major component of computer assisted translation, but how does a translator actually access terminology data bases ? Three main categories emerge from the way the market has been developing more powerful and less expensive hardware, and corporate software applications, usually in a Windows environment, sometimes in connection with European initiatives (Genelex, Eurolang...).

The first category is made up of classical packages, still well represented, but by now decidedly down-market. There are no more than there were, as new products simply replace those that disappear. They are many and varied as to form : some boast few functions, others are much more fast, powerful and user friendly.

Other software packages are designed as part of a complete translation work station, integrating translation memories, terminology management and various automatic functions.

The third category is made up of primarily lexicographical products, designed "generically", and meant to be used in the most varied situations, including as dictionaries built into automatic translation systems and by the same token, in terminology.

It remains to be seen what actual use can be made of these products in the field of translation. Who are they targeted at ? Are they really adapted to translators needs ?

Introduction

Le Centre de terminologie et de néologie du Centre national de la recherche scientifique, à Paris, est avant tout un centre de documentation sur la terminologie, ce qui implique, entre autres tâches, celle d'identifier et de recenser, non seulement les produits terminologiques existants, comme les dictionnaires ou les ouvrages théoriques, mais aussi les outils d'aide à la constitution de terminologie : les logiciels.

Les logiciels de gestion de données terminologiques, à l'instar de la terminologie elle-même, se situent généralement dans un contexte de traduction, et se présentent même parfois comme des outils exclusivement dédiés aux traducteurs et totalement adaptés à leurs besoins. Cette affirmation paraît assez optimiste, pour au moins deux raisons.

D'une part il y a traducteur et traducteur. Qu'y a-t-il de commun entre les besoins du traducteur indépendant et ceux d'un traducteur qui travaille dans le cadre du service de traduction d'une grosse entreprise ? D'autre part, les logiciels semblent souvent le résultat non pas d'une coopération entre un concepteur et un traducteur mais plutôt d'une prise de pouvoir du concepteur, qui néglige l'importance de l'expérience du traducteur. Il estime, au bout d'une conversation d'une heure, avoir tout compris de tous les besoins de tous les traducteurs. À moins que ce ne soit le commercial qui surestime le produit qu'il doit vendre. On constate en tout cas encore un fossé entre ce que les traducteurs attendent et ce dont ils disposent.

Heureusement le marché est en pleine évolution, en réelle évolution, grâce à des mutations dans le marché général de l'informatique, et grâce à un renouvellement de l'offre logicielle qui bénéficie des retombées des programmes de recherches dans ce domaine.

Évolution du marché de l'informatique

Ce qui suit n'est une surprise pour personne, mais il convient de le souligner comme préalable. L'unanimité s'est faite, ces dernières années, autour de deux grands types de produits : les ordinateurs compatibles avec les PC d'IBM, d'une part, et les Macintosh, d'autre part, bien qu'en moindre proportion peut-être dans nos professions. Windows a désormais acquis un statut de quasi standard dans le monde PC. De plus le prix des ordinateurs (surtout ou d'abord les PC) s'est effondré en même temps que leur puissance augmentait.

Quelles sont les conséquences de cette évolution du marché ?

Tout un chacun a désormais accès à des ordinateurs puissants, rapides, bénéficiant d'une capacité de stockage importante. Les développeurs mettent à profit cette nouvelle liberté pour concevoir des programmes plus faciles à utiliser (ou du moins pourraient la mettre à profit). La convivialité est à l'ordre du jour avec Windows (même si nous avons quelques réserves). Bien que chacun puisse citer des exemples célèbres de traducteurs ou de terminologues qui le refusent, l'ordinateur est désormais présent dans chaque bureau et presque dans chaque foyer. Cette vulgarisation accrue de l'outil informatique correspond à un changement dans la manière dont il est perçu. Les utilisateurs n'ont plus peur (ou ont moins peur) des *informagiciens*, pour reprendre le mot de Maryvonne Abraham. Ils ne sont pas pour autant plus compétents en informatique, mais, plus conscients des possibilités qu'offrent l'informatique, ils sont devenus plus exigeants. Ils veulent des programmes simples, conviviaux, transparents, avec souris, icônes et menus déroulants, aide en ligne etc. Et c'est possible. Mais les logiciels de terminologie ont-ils vraiment subi la même évolution que leur public ?

Les logiciels de terminologie

Traducteurs et terminologues n'ont que l'embarras du choix. On compte en effet actuellement une quarantaine de produits différents sur le marché, toutes catégories confondues, commercialisés, en cours de développement ou encore logiciels dits maison. Hélas, il faut bien reconnaître que nombre d'entre eux ne font que confirmer ce que Jean-Michel Henning disait déjà en 1985, à savoir qu'ils correspondent à un état relativement ancien de la technologie. On continue à proposer des produits qui ne sont en fait que de petites programmations de logiciels de gestion de base de données. Heureusement certains se tournent vers des outils différents, issus de l'intelligence artificielle, pour proposer des produits radicalement nouveaux.

On distinguait autrefois les logiciels pour traducteurs et les logiciels pour terminologues : les premiers directement accessibles à partir d'un logiciel de traitement de texte, les autres fonctionnant – ou pouvant fonctionner – en mode autonome. Avec Windows cette différence n'a plus lieu d'être, si ce n'est qu'il reste quand même plus facile – plus rapide – de faire du couper-coller entre des logiciels conçus pour Windows.

Un autre clivage pourrait se faire en fonction du type de matériel utilisé. Mais outre le fait que le choix d'un logiciel se fait plutôt après celui du matériel, il faut bien dire que les PC se taillent là encore la part du lion, sous DOS ou sous Windows. Il n'existe à notre connaissance qu'un seul produit pour OS2, les logiciels pour Mac sont minoritaires, et Unix n'est le support privilégié que de quelques nouveaux logiciels haut de gamme.

Quel peut donc être le critère de classement de ce type de logiciels ? Signalons au passage que cette question sera précisément l'un des thèmes de réflexion du tout nouveau groupe de travail 11 de l'Association pour la terminologie et le transfert des connaissances (GTW selon ses initiales allemandes).

On peut en tout cas distinguer trois tendances majeures : les logiciels classiques, qui sont nés au début des années 80, les logiciels *tout en un* ou *intégrés*, et des produits très élaborés et destinés à la terminologie ou la lexicographie très fines ou même à la recherche dans ces domaines.

Les logiciels classiques

Il faut encore opérer des subdivisions dans cette catégorie. Certains logiciels semblent avoir totalement disparu, d'autres n'ont pas du tout évolué depuis plusieurs années, d'autres encore se sont adaptés. Certains logiciels ne tournent pas encore sous Windows, d'autres l'acceptent, d'autres encore sont conçus pour Windows. Certains logiciels se suffisent à eux-mêmes, d'autres nécessitent l'adjonction d'un autre logiciel, généralement un SGBD. Certains n'autorisent qu'une description terminologique des plus succinctes (terme, équivalent, plus un très court champ description), d'autres laissent toute liberté au terminologue, d'autres encore lui laissent une très large marge de manœuvre. Parmi les nouveaux logiciels du marché, il faut aussi distinguer ceux qui sont de structure classique et ceux qui représentent une nouvelle tendance.

La plupart de ces logiciels ont déjà fait l'objet d'études approfondies qu'il est hors de question de reproduire ici. Mentionnons simplement la disparition vraisemblable de Term-Lidas et d'Omniterm, sur lesquels il est impossible d'obtenir la moindre information, de Profilex, dont le concepteur est décédé, et la disparition avérée de Dicoterm, superbe logiciel victime des commerciaux.

Un petit mot d'un logiciel qui a le vent en poupe actuellement : Multiterm, développé par la société allemande Trados. Multiterm existe en deux versions, une pour DOS et une pour Windows. La version pour DOS est également commercialisée sous le nom de Termtracer par l'ex-société INK, devenue RR Donnelley Language Solution depuis son rachat par cette société américaine. Il faut aussi préciser que Multiterm peut être utilisé comme module de terminologie d'un poste de travail complet, commercialisé sous le nom de Translator's Workbench. Mais il faut se garder de confondre ce Translator's Workbench avec le projet européen de même nom. Précisons encore que des accords commerciaux entre Multiterm et Termex vont permettre aux utilisateurs de Multiterm d'avoir accès à tous les glossaires déjà disponibles pour Termex.

Parmi les nouveaux logiciels, on l'a déjà mentionné, certains sont de structure classique et sont souvent développés sur Clipper ou sur un noyau de SGBD comme dBase. Dans cette catégorie on peut mentionner PROTERM (les logiciels Tradulog), déjà bien connu ici, ou, plus récents, NÉOLOG (nom provisoire) développé par le Centre de terminologie de l'École de droit de l'Université de Moncton, LEXPRO (nom provisoire) par la société française LCI, AUTOLEX, par le Centre de terminologie et de traduction spécialisées de l'Académie des sciences de Cuba, FAOTERM par le Service de terminologie de la FAO à Rome, WHOTERM, par le Service de terminologie de l'Organisation mondiale de la santé à Genève, LEXIKON par le Service national de terminologie d'Afrique du Sud, TERMYSIS, par la petite société allemande Köller, CONCEPT & TERM par l'École des hautes études commerciales de Copenhague, ou encore quelques autres. Il faudrait y ajouter TERMISTI développé par l'ISTI de Bruxelles, qui a l'avantage de permettre un contrôle et une structuration du réseau notionnel (voir Van Campenhoudt, dans cet ouvrage). Précisons encore que tous n'en sont pas au même degré d'achèvement ou de commercialisation.

Une autre catégorie de nouveaux logiciels fait la part belle au dépouillement assisté par ordinateur et à la représentation des relations notionnelles sous forme de graphe. Pour la plupart ils utilisent des outils issus des recherches en intelligence artificielle. Ce type de logiciel correspond peut-être plus aux besoins des terminologues qu'à ceux des traducteurs, mais ils peuvent être très utiles dans le cadre d'un service de traduction-terminologie. COGNITERM (voir Meyer, dans cet ouvrage) met l'accent sur une structuration dynamique du domaine, qui permet au terminologue de ne perdre aucune information. HYPERTEPA est très proche de Cogniterm par sa conception : un système hypertexte permet à l'utilisateur d'aller d'un enregistrement à un autre, ou au réseau notionnel, représenté sous forme de graphique, et de repartir du réseau vers un enregistrement. Ce logiciel est développé par le Centre finlandais de terminologie technique (TSK).

Parmi les logiciels peut-être plus directement utilisables par les traducteurs, il faut citer SYSTEM QUIRK (autrefois appelé MATE), développé par le Groupe d'intelligence artificielle de l'Université du Surrey. Ce logiciel a été mis au point dans le

cadre des projets européens Translator's Workbench et Multilex et se situe à mi-chemin des logiciels de terminologie pour traducteurs et des logiciels intégrés. En effet il inclut des procédures de repérage automatique de termes (module KonText), de gestion de corpus (Corpus Manager), d'édition selon différents formats prédéfinis (Lexicon Publisher), et surtout de gestion de réseau notionnel (Word Linker), avec affichage graphique des liens entre notions. System Quirk est actuellement utilisé, semble-t-il, par le service de traduction de l'Université du Surrey, par la Communauté européenne et par la société française SITE, partie prenante du projet Multilex.

Les logiciels intégrés

Ces nouveaux logiciels se sont fixé pour but de soulager le traducteur dans le cas de traductions très répétitives comme les versions successives d'un même logiciel ou d'un système de télécommunication.

Ils fonctionnent sur le principe de la mémoire de traduction : les traductions antérieures, qui peuvent être regroupées par l'utilisateur en sous-ensemble ou dossiers sur un même thème, servent de base de référence pour traduire les nouveaux textes.

Le logiciel compare alors le nouveau texte source à cette base de référence, identifie les segments qui ont déjà fait l'objet d'une traduction, et propose donc la traduction correspondante, que l'utilisateur peut alors réutiliser avec ou sans modification ou ignorer. On parle de segment et non de phrase car l'utilisateur peut définir des segments de la longueur qu'il souhaite, des paragraphes par exemple, la seule contrainte étant qu'il doit y avoir le même nombre de segments dans le texte source et dans le texte cible (ce qui peut poser un certain nombre de problèmes).

De même l'utilisateur peut dans certains cas définir le degré de ressemblance qu'il souhaite entre les segments : identité totale, pourcentage de mots pouvant varier, types d'éléments dont la variation ne sera pas prise en compte (chiffres par exemple), etc.

Dans tous les cas, le logiciel procède à une analyse préalable du texte source, qui doit, cela va sans dire, être disponible sur support électronique.

Ce type de logiciel comporte généralement un module de gestion terminologique – que l'on peut définir comme étant la plupart du temps assez limité dans la description qu'il permet des termes. Ce module terminologique signale à l'utilisateur les termes déjà connus, ou, dans certains cas et selon le choix de l'utilisateur, les termes absents du dictionnaire.

Ces logiciels incorporent souvent des possibilités d'édition et de mise en page au moins compatible avec des logiciels de PAO et parfois directement au format de ces logiciels.

Il eût été possible de classer System Quirk dans cette catégorie de logiciel car il répond parfaitement à cette description. Son module de terminologie très développé avec l'accent mis sur la mise en évidence des relations notionnelles et la visualisation des réseaux correspondants en font toutefois un logiciel à part et beaucoup plus complet que les autres à cet égard.

Parmi les systèmes effectivement commercialisés, on peut mentionner TM2 (Translation Manager) qu'IBM a lancé en juillet 1992. Ce logiciel est le seul de cette catégorie, à ma connaissance tout au moins, à tourner sous OS2.

TRANSIT, commercialisé par la société allemande Star, est un produit très similaire à TM2.

Mentionnons encore rapidement un produit qui n'est pas encore disponible mais dont la version prototype, utilisée chez SITE (France) est très intéressante par la souplesse qu'elle offre : il s'agit du logiciel mis au point dans le cadre du projet européen EUROLANG.

Les logiciels de recherche

Il convient de signaler ces produits, qui, bien qu'ils aillent très au-delà des attentes – et des possibilités matérielles actuelles – des traducteurs, représentent une étape intéressante et prometteuse dans l'évolution du marché.

Dans le cadre du projet européen GENELEX, on attend la commercialisation de logiciels de description lexicale très fouillée, tant au niveau morphologique que syntaxique et sémantique, et qui, du fait de l'adoption du format SGML, pourront être réexploités à diverses fins, dans la logique du projet qui souhaitait viser la généricité.

Le Lexicaliste est un produit déjà disponible, commercialisé par la société SITE, développé par elle pour le Centre national d'étude des télécommunications. Ce logiciel qui tourne lui aussi sur gros système et qui lui aussi fournit des données au format SGML, propose une description très poussée des entrées lexicales, avec en particulier la possibilité d'établir des liens, lexicaux ou sémantiques, avec la génération des liens inverses et tous les contrôles de cohérence possibles, ainsi que la possibilité pour l'utilisateur de définir son environnement de travail comme il le souhaite.

Conclusion

Dans l'état actuel des choses, ne vaut-il pas mieux privilégier un outil simple mais qu'on connaît bien, plutôt qu'un logiciel très sophistiqué qui va peut-être faire perdre plus de temps qu'il n'en fera gagner ?

C'est en tout cas l'avis des utilisateurs. Un rapide sondage m'a en effet permis de constater qu'apparemment la plupart des traducteurs français, qu'ils soient indépendants ou en entreprise, n'utilisent pas de logiciel de terminologie : un tiers utilise le bon vieux traitement de texte, un autre tiers un logiciel de gestion de données comme dBase 3 ou 4. Il semblerait qu'au Québec la situation soit assez semblable, avec une utilisation prédominante de logiciels comme Filemaker ou Edibase. Qu'en est-il aujourd'hui et surtout qu'en sera-t-il demain ?

Une certitude en tout cas : les nouveaux outils qui apparaissent sur le marché sont extrêmement prometteurs. Certes ils n'ont pas encore tous été testés en situation

de travail par des traducteurs, et certains ne sont même pas encore commercialisés, mais il paraît évident qu'ils devraient permettre une grande économie de temps et d'énergie aux traducteurs.

Le problème majeur est à notre avis le manque d'information : les utilisateurs potentiels sont effrayés par l'abondance des offres. Ils ont parfois aussi eu dans le passé des expériences assez désagréables et sont devenus très méfiants.

Ce que le marché leur propose désormais, ou leur proposera bientôt, semble quand même beaucoup mieux correspondre à leurs besoins réels. Reste qu'il doivent s'en convaincre eux-mêmes.

Références

ABRAHAM, Maryvonne (à paraître) : « Une messagerie d'informagiciens », *Quatrièmes journées ERLA-GLAT*, UBO-ENST Bretagne, Brest.

AUGER, Pierre (1991) : « Terminographie et lexicographie assistées par ordinateur : état de la situation et prospectives », *Actes du colloque Les industries de la langue, perspectives des années 1990,* Montréal 21-24 novembre 1990, Québec, OLF-STQ.

BÉDARD, Claude (1990) : « Quoi de neuf en traductique », *Circuit,* n° 30, septembre, Montréal, Société des traducteurs du Québec, pp. Q3-Q14.

BELL, Sabine (1992) : « Transit, the Ideal Working Environment for Translators », *Language International,* 4-6, Londres.

DE BESSÉ, Bruno et Donatella PULITANO (1990) : « Les logiciels de gestion de la terminologie », *Terminologie et Traduction,* n° 3.

BLANCHON, Élisabeth (1992) : « Choisir un logiciel de terminologie », *La Banque des Mots.* Numéro spécial, Paris, CTN 1991, CILF.

BLANCHON, Élisabeth (1992) : « Comparaison de logiciels utilisables en terminologie », A. Clas et H. Safar (dir), *L'environnement traductionnel. La station de travail du traducteur de l'an 2001,* Actes du Colloque de Mons (Belgique), Mons, 25-27 avril 1991, actualité scientifique, Sillery, AUPELF-UREF, Presses de l'Université du Québec, pp. 223-233.

BURDET, Claude-Alain (1991) : « Terminology and the Management of Information, some Practical Solutions Delivered by the Notional Inference Engine of Termis », *Actes du symposium international Terminologie et documentation dans la communication spécialisée,* Montréal, Infoterm, Secrétariat d'État du Canada.

FREIGANG, Karl-Hanz, MAYER, Félix et Klaus-Dirk SCHMITZ (1991) : « Micro- and Minicomputer-based Terminology Databases in Europe », *TermNet Report* 1991.

HENNING, Jean-Michel (1986) : « L'évolution des logiciels de gestion terminologique », *Actes du colloque Terminologie et technologies nouvelles,* Office de la langue française et Conseil général de la langue française, Paris 9-11 décembre 1985, Québec, OLF.

MAURICE, Nathalie (1991) : *Évaluation des logiciels MC4, Aquila et Foxbase + dans une perspective terminologique,* mémoire présenté pour le DESS « Information et documentation », Institut d'études politiques de Paris, 1989, Publié aux Éditions de l'ADBS, Paris.

MAURICE, Nathalie, BLANCHON, Élisabeth, OTMAN, Gabriel et Jacques BOISSY (1991) : « Comparaison de trois logiciels utilisables en terminologie: Foxbase +, MC4, Texto », *Meta*, vol. 36, nº 1, mars 1991.

MAURICE, Nathalie (1990) : « Conception d'une base de données terminologiques multilingue dans le domaine du droit, Analyse des besoins et méthode suivie », *Terminologie et Traduction*, nº 3.

MAYER, Felix (1990) : « Terminologieverwaltungssysteme für Übersetzer », *Lebende Sprachen*, nº 3.

MAYER, Felix (1990) : « Export und Import von Daten bei Terminologieverwaltungssytemen », *Terminologie et Traduction*, nº 3.

MEYER, Ingrid, BOWKER, Lynne et Karen ECK (1992) : « COGNITERM : An Experiment in Building a Terminological Knowledge Base », *Proceedings of the Fifth Euralex International Congress*, Tampere, Finland, 4-9 août 1992.

MEYER, Ingrid, BOWKER, Lynne et Karen ECK (1991) : « Constructing a Knowledge-Based Term Bank: Fundamentals and Implications », *Actes du symposium international Terminologie et documentation dans la communication spécialisée*, Infoterm, Montréal, Secrétariat d'État du Canada.

DE SCHAETZEN, Caroline (1990) : « Bilan des dictionnaires électroniques et gestionnaires de dictionnaires », *Terminologie et Traduction*, nº 3.

DE SCHAETZEN, Caroline (1990) : « L'ordinateur peut-il fabriquer des dictionnaires », *La Banque des Mots*, nº 40.

DE SCHAETZEN, Caroline (1990) : « Outils de bureautique et de télématique pour la traduction », *Lebende Sprachen*, nº 3.

SCHMITZ, Klaus-Dirk (1990) : « Rechnergestützte Terminologie-verwaltung am Übersetzerarbeitplatz », *Terminologie et Traduction*, nº 3.

SITE (1992) : *Le Lexicaliste, Manuel de référence, Manuel d'utilisation.*

VERHAEST, Frank (1991) : « Évaluation de cinq gestionnaires de glossaires », *Actes du colloque Les industries de la langue. Perspectives des années 1990*, Montréal 21-24 novembre 1990, Québec, Office de la langue française et Société des traducteurs du Québec.

Répertoire des produits et services de traitement automatique de la langue française (1989) : Observatoire des industries de la langue, Paris, Daicadif.

Terminogramme (1990) : Québec, Office de la langue française, nº 55, hiver 1990.

30

Terminologie à l'Union de Banques Suisses

Patrick BURKHARD

Union de Banques Suisses, Zürich, Suisse

• *Abstract* •

UBS commissioned Digital Equipment Corp. (DEC) to develop a new terminology management software package to meet the specific needs of end users. DEC will market the product under the name TerMS©. Its structure, which will be briefly described, strongly influences the way terminologists work:
– the terminology cards created are monolingual but can be interrogated in a multilingual mode ;
– the user groups are very diverse and expect different types of information (linguistic, encyclopedic, contextual).

The last part of this presentation will focus on the perspectives offered by this type of tool not only in the field of computer-integrated translation, but also in a modern office environment.

Introduction

L'Union de Banques Suisses (UBS) est le premier groupe bancaire de Suisse. Son activité s'inscrit dans un paysage linguistique où cohabitent l'allemand, le français, l'italien et le romanche – seules les trois premières ont le statut de langue officielle. Par ailleurs, la solide implantation de l'UBS à l'étranger et sur les marchés internationaux donne une importance accrue à l'anglais qui tend à devenir la langue officielle du groupe.

Cette pluralité des langues de travail entrave la communication au sein de l'entreprise et avec la clientèle. Elle rend le maintien et l'usage d'une terminologie d'entreprise consistante particulièrement ardus. Pour remédier à ces difficultés, l'UBS a équipé son service de traduction :

- d'un système de TA pour la traduction allemand-anglais (METAL) ;
- d'une banque de données terminologiques quadrilingue.

Elle s'est ainsi donné les moyens d'offrir des produits d'une qualité linguistique optimale à sa clientèle et à ses collaborateurs, valorisant ainsi l'activité du traducteur.

Quelle banque de terminologie ?

Pour sa future banque de terminologie, l'UBS cherchait un produit répondant à des exigences bien précises.

Le futur système devait permettre une large diffusion du fonds terminologique, de sorte que tout collaborateur de la banque dont l'activité est liée de près ou de loin à la langue (rédacteur technique, documentaliste, secrétaire, stagiaire etc.) puisse profiter des recherches du service de traduction.

Afin de disposer d'un système aussi ouvert que possible, le logiciel devait pouvoir gérer un nombre illimité de langues susceptibles d'être à la fois langue source et langue cible. La structure de la banque de données devait également permettre le couplage avec d'autres logiciels moyennant un minimum d'efforts. De plus, le système devait s'appuyer sur des techniques d'avant-garde afin d'être en mesure d'offrir une convivialité maximale à tous les utilisateurs.

Un nouveau logiciel : TerMS[1]

L'UBS a confié le développement d'un logiciel à la société Digital. La nouvelle banque de données terminologiques **TerMS** (acronyme de Terminology Management System) sera commercialisée par Digital une fois le développement terminé.

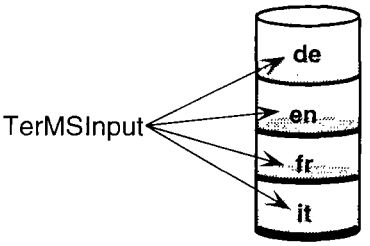
Structure modulaire

Pour que chaque langue puisse remplir à la fois la fonction de langue source et celle de langue cible, les ingénieurs ont proposé de créer une banque de données par langue.

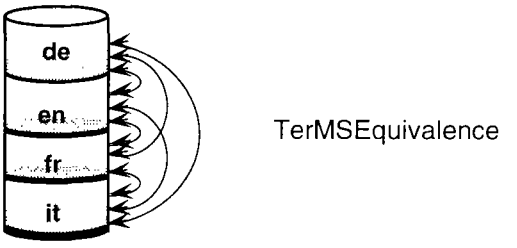
Le logiciel est constitué de quatre modules :

a) TerMSInput est le module de saisie des fiches terminologiques. Il permet la création et le maintien d'un fonds unilingue pour chacune des langues de travail.

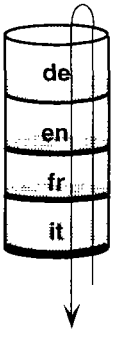
1. Digital Equipment Corporation est propriétaire de l'appellation TerMS.



b) TerMSEquivalence est le module utilisé pour exprimer les relations d'équivalence entre les entrées des différentes langues de travail.



c) TerMSLookup est le module d'interrogation.



d) TerMSDatabaseManagement est le module de gestion de la banque de données. Il permet à l'administrateur de gérer la banque de données sans passer par le système d'exploitation. Les principales opérations possibles sont :

- la gestion de tables dynamiques, par exemple la table des domaines ;
- la gestion de l'accès aux différents modules et banques de données ;
- la gestion des échanges de terminologie (import/export).

Format terminologique

Pour éviter les champs « fourre-tout » et permettre une interrogation ciblée de la banque de terminologie, nous avons voulu séparer chaque type d'information. Par conséquent, le nombre de champs de saisie est assez élevé, mais cela n'affecte en rien l'interrogation, les champs vides n'étant pas affichés à l'écran.

Champs de saisie et contenu (les champs obligatoires sont indiqués en gras) :

Domaine :	Indication du domaine auquel la fiche est assignée.
Terme :	Vedette de la fiche – avec indication grammaticale.
Note linguistique :	Information complémentaire de type métalinguistique sur l'usage du terme, d'un synonyme, d'une abréviation ou d'une variante géographique.
Source terme :	Source du terme, le cas échéant d'un synonyme, d'une abréviation, d'une variante ou de la note linguistique.
Synonyme :	Toute désignation du concept qui n'est ni la vedette, ni une forme abrégée, ni une variante géographique – avec indication grammaticale.
Abréviation :	Toute forme abrégée du terme – avec indication grammaticale.
Variante :	Variante géographique du terme, c'est-à-dire étrangère à l'usage dans la zone linguistique de référence – avec indication grammaticale. À l'UBS : usage suisse (pour l'allemand, le français et l'italien) et américain (pour l'anglais).
Définition :	Description synthétique du concept.
Source :	Source de la définition, le cas échéant de la note.
Note :	Information encyclopédique ne trouvant pas sa place dans une définition.
Renvoi :	Terme associé, par la forme ou le contenu au concept défini, avec possibilité d'indiquer si le terme en question est défini dans TerMS.
Exemple :	Attestation de l'usage du terme en contexte – avec source.
Collocation :	Construction usuelle (adjective ou verbale), figée ou non – avec note d'usage.
Cote :	Fiabilité du contenu : « en chantier », non approuvé, approuvé ; accessibilité de la fiche : possibilité d'en restreindre l'accès à un groupe d'utilisateurs.
Données administratives :	Champs saisis automatiquement par TerMS – auteur et date de création, réviseur et date de révision, provenance et numéro de la fiche.

En outre, les informations suivantes peuvent être consultées lors de l'interrogation :

Traduction :	Équivalence de la vedette dans la (les) langue(s) choisie(s).
Traduction des collocations :	Construction usuelle équivalente du terme dans la (les) langue(s) choisie(s).

Convivialité

La convivialité du système est un objectif que nous n'avons jamais perdu de vue pendant le développement de TerMS.

Dans TerMSLookup, les valeurs par défaut du système permettent l'interrogation sans détour de la banque de terminologie. Lors de recherches plus pointues, l'utilisateur peut également choisir et enregistrer des paramètres personnalisés.

Dans TerMSInput, convivialité signifie en particulier longueur des champs de saisie illimitée et saisie sans codes ni abréviations.

Terminologie à l'UBS

Public cible hétérogène

Le but du service de terminologie de l'UBS est d'atteindre et de satisfaire un public très vaste, allant du traducteur à la secrétaire en passant par le rédacteur technique, l'apprenti et le collaborateur spécialisé. Leurs attentes quant au contenu de la banque de terminologie ne sont pas les mêmes. Certains sont plus intéressés par le contenu encyclopédique de la fiche (définition et note), d'autres uniquement par les traductions du terme et les collocations, d'autres encore par le contenu linguistique (terme, synonymes, abréviations, variantes) ou phraséologique (exemples et collocations).

Par conséquent, il faut que chaque utilisateur ait la possibilité de sélectionner les paramètres de recherche qui lui permettront de trouver l'information le plus simplement possible. Il est également important que chacun puisse choisir un format d'édition adapté à ses besoins, afin de ne pas être submergé d'informations superflues. Bref, le module d'interrogation du fonds de terminologie doit être flexible.

C'est en réponse à cet impératif que nous avons jugé nécessaire de multiplier les champs de saisie. En effet, il faut trier l'information dès le départ, c'est-à-dire dès la saisie, pour que l'utilisateur dispose de paramètres de recherche et de sortie précis lors de l'interrogation.

Approche unilingue

L'approche unilingue imposée par la conception de TerMS est certainement positive dans un cas comme le nôtre, où le nombre de champs de saisie est relativement élevé. Le terminologue a ainsi la possibilité de se concentrer sur le contenu, puis sur les équivalences. Ceci est particulièrement vrai lors d'un travail thématique.

La conséquence principale est que le concept de langue source et langue cible n'existe plus, au moins lors de la saisie, et qu'ainsi aucune des langues n'est privilégiée.

Pour obtenir une qualité de fiches optimale, nos terminologues, qui sont égale-

ment traducteurs, travaillent presque tous dans une seule langue. Après avoir saisi les fiches, ils établissent les équivalences en équipe.

Nous trouvons cette méthode avantageuse. Elle permet à plusieurs terminologues de couvrir le même sujet et de comparer ensuite le résultat de leurs recherches. Nous pouvons ainsi garantir une qualité de contenu égale pour toutes les langues, ce qu'il est parfois difficile de réaliser sur une base multilingue.

Pour les recherches ponctuelles nous travaillons généralement sur une base bilingue avec l'allemand comme langue de départ, ou à tout le moins comme langue de référence, car le siège du groupe se trouve à Zurich, dans la partie germanophone du pays. La plus grande partie des textes sont donc écrits en allemand, raison pour laquelle la demande de traduction concerne principalement des paires de langues comprenant l'allemand.

Modes d'utilisation du fonds de terminologie

Utilisation directe

L'utilisation usuelle et aussi la plus fréquente du fonds de terminologie est l'interrogation au moyen du module d'interrogation TerMSLookup. Celui-ci offre différentes possibilités de varier et de personnaliser l'interrogation :

- paramètres d'interrogation personnalisés ;
- création de formats d'édition personnalisés ;
- extraction de sous-ensembles de la banque de données ;
- transfert direct de l'information dans un logiciel de traitement de texte ;
- impression.

Utilisation indirecte

Le but de ce type d'utilisation est la normalisation de la terminologie d'entreprise par l'intermédiaire d'autres systèmes informatisés que le logiciel de gestion du fonds de terminologie. En d'autres termes, il s'agit de permettre une utilisation consistante de la terminologie d'entreprise sans devoir obligatoirement passer par TerMS. Il est à noter qu'aucune des propositions présentées ci-après n'a été réalisée pour l'instant. Cependant elles sont prévues depuis le début du projet et l'architecture de TerMS permet de les réaliser avec un minimum d'efforts de développement.

Échange avec un système de TA/TAO

L'échange régulier de terminologie entre le système de TA/TAO et la banque de données terminologiques permettra d'éviter des disparités entre la terminologie utilisée par les traducteurs d'une part et le système de TA/TAO de l'autre. Les informations échangées seront le terme, les équivalences, les indications grammaticales (de base) et le domaine.

Utilisation par l'intermédiaire d'autres logiciels

Intégration dans les applications de l'entreprise sous la forme d'aide en ligne.

Intégration, voire fusion avec le correcteur orthographique des systèmes de traitement de texte utilisés dans l'entreprise.

Utilisation en tant que thésaurus, pour le service de documentation ou à des fins de rappel de textes (*text retrieval*).

Conclusion

Le concept sur lequel repose la banque de terminologie TerMS contraint le terminologue de travailler sur une base unilingue, ce qui, pour nous constitue un avantage permettant de garantir la qualité des fiches. Il en va de même pour le nombre de champs à disposition qui permet de développer à la fois les aspects linguistiques et encyclopédiques de l'entrée et du concept.

Par ailleurs, TerMS offre des perspectives nouvelles d'utilisation du fonds terminologique. Il contient des informations diversifiées et élargit ainsi le cercle des utilisateurs. Pour cette raison, il est possible d'envisager une utilisation moins conventionnelle du logiciel et d'intégrer le fonds de terminologie à la bureautique de l'entreprise, que ce soit sous la forme de « dictionnaire électronique », ou, comme nous l'avons vu plus haut, pour compléter des vérificateurs orthographiques et des aides en ligne.

31

Principes directeurs pour l'établissement d'une banque des morphèmes-racines de l'arabe standard

Hussein HABAILI

Université de Tunis I et Institut régional des sciences informatiques et télécommunications à distance, Tunis, Tunisie

• *Abstract* •

This paper presents the guidelines and the principles for the establishment of a Databank of standard Arabic root morphemes. It describes the development of the Arabic lexicon and its subsequent testing as a Databank which includes:
– a theoretical lexicon of triradical Arabic root morphemes ;
– an attested lexicon, of Arabic triradical root morphemes ;
– a non attested lexicon, but admissible triradical root morphemes of standard Arabic.

Our Databank of Arabic triradical root morphemes employs a computational technique which allows for a given root morpheme to be tested as:
– an attested root morpheme ;
– or, as a non attested but admissible root according to the morpheme structure conditions of standard Arabic.

Introduction

La production et l'élaboration de dictionnaires électroniques et de *Banques de morphèmes-racines* sont devenues l'objet privilégié de la communication **homme-machine**. En effet, ces activités sont considérées comme les domaines de recherche de tout premier plan en *intelligence artificielle* et en *linguistique computationnelle.*

Les *industries de la langue* sont naissantes, cependant les études menées ces dernières années dans divers pays, et plus particulièrement en France et au Canada, montrent que ces industries sont appelées à un essor considérable.

Notre article s'inscrit dans le cadre d'un projet de recherche mené à l'*Institut régional des sciences informatiques et télécommunications à distance (IRSIT)* de Tunis.

Ce projet a pour objet l'établissement d'une banque des morphèmes-racines (trilitères d'abord, bilitères et quadrilitères ensuite) de l'arabe standard.

Ce travail présentera les principes directeurs d'ordre linguistique et informatique adoptés lors de l'établissement de notre banque des racines arabes. Cette banque sera à la base de toute analyse linguistique dans différents domaines :

- l'élaboration de dictionnaires électroniques ;
- la traduction automatique ;
- la traduction assistée par ordinateur ;
- la synthèse de la parole arabe ;
- la conjugaison automatique ;
- la correction orthographique ;
- l'analyse morphologique automatique ;
- la néologie lexicale...

Nos principes linguistiques sont basés sur la notion de conditions de structure morphématique de l'arabe, élaborée par Habaili (1990). Les conditions permettent de délimiter l'ensemble des matrices phonologiques possibles qui forment les racines admissibles de l'arabe standard.

Notre approche computationnelle utilise la méthode SADT qui est semi-formelle (seule la syntaxe est formelle alors que la sémantique est informelle), descendante, modulaire, hiérarchique, structurée et graphique. Cette méthode permet une description fonctionnelle du système, indépendante des diverses solutions envisageables pour sa réalisation. Elle utilise la notion de diagramme, d'actigramme et la conception fonctionnelle descendante.

Principes linguistiques

Sur le plan lexicologique et linguistique, les principes sont de deux types :

– Ils sont basés, d'une part, sur la notion de « conditions de structure morphématique » (CSM), en ce qui concerne l'élaboration des différents types de dictionnaires (théorique, attesté, admissible, non admissible) de l'arabe standard.

– Sur le plan lexicographique, nos principes s'appuient sur les grands dictionnaires arabes, tels que le dictionnaire *As-sihah* d'Al Gawhari ; *Lisân 'Al 'Arab* d'Ibn Mandhûr ; et *Tâj 'Al 'Arûs* d'Az-Zubeidî, qui sont la base de notre grand dictionnaire attesté de l'arabe standard.

Les conditions de structure morphématique de l'arabe standard

Les conditions de structure morphématique sont des contraintes profondes qui permettent de délimiter l'ensemble des matrices lexicales et phonologiques possibles qui

forment les morphèmes-racines admissibles de l'arabe standard.

On distingue deux types de conditions de structure morphématique :

– Les *restrictions combinatoires* qui lient les spécifications des différents traits à l'intérieur d'un même segment de la matrice phonologique ;

– Les *restrictions séquentielles* qui lient les spécifications de traits appartenant à des segments successifs d'une même matrice, ce qui implique que n'importe quelle séquence de phonèmes de l'arabe standard n'est pas nécessairement un morphème-racine.

Le modèle chomskyen (1968) et le formalisme de Stanley (1967) nous ont fourni quatre types de conditions de structure morphématique, à savoir : les *conditions positives, négatives, si-alors , si et si-alors.* Nous pouvons donc interpréter ces conditions comme spécifiant les coefficients de traits particuliers dans des contextes particuliers. Il est donc naturel de proposer que ces conditions soient incorporées à la grammaire de l'arabe standard et que les traits prédictibles demeurent non spécifiés dans les entrées lexicales.

Dans un travail antérieur (Habaili 1990), nous avons élaboré, pour l'arabe standard, cinq conditions de structure morphématique qui rendent compte de toutes les combinaisons possibles, admissibles et non admissibles des racines trilitères de l'arabe standard.

La première condition CSM_1 est une condition négative qui rend compte de toutes les combinaisons possibles entre tous les segments non syllabiques de l'arabe standard en position 1,2 :

$$CSM_1 \quad + \quad \tilde{} [\alpha \text{ Traits}] \quad [\alpha \text{ Traits}]$$

En effet, la condition négative CSM_1 exclut d'une façon absolue toute combinaison de segments non syllabiques identiques en position 1,2.

En ce qui concerne les segments homorganiques non identiques, nous pouvons avoir la condition *si alors* suivante :

$$\begin{array}{lllll} & & & [\alpha \text{ ant}] & [\alpha \text{ ant}] \\ & & & [\beta \text{ cor}] & [\beta \text{ cor}] \\ CSM_2 : & \text{Si} & + & [\gamma \text{haut}] & [\gamma \text{haut}] \\ & & & [\delta \text{ bas}] & [\delta \text{ bas}] \\ & & & [\epsilon \text{ arr}] & [\epsilon \text{ arr}] \\ & & & & \\ & \text{alors} & + & [\zeta \text{ son}] & [-\zeta \text{ son}] \\ & & & [\eta \text{ cont}] & [-\eta \text{ cont}] \end{array}$$

Cette condition CSM_2 accepte toute combinaison de segments homorganiques non identiques qui satisfont la condition *alors,* c'est-à-dire toute combinaison de deux segments dont l'un diffère de l'autre, soit par le son ou en continuité.

En position 2,3, nous pouvons avoir la condition suivante :

CSM_3 :	Si	[α Traits]	[α Traits]	+
	et si	[α ant]	[α ant]	
		[β cor]	[β cor]	
		[γ haut]	[γ haut]	+
		[δ bas]	[δ bas]	
		[ϵ arr]	[ϵ arr]	
	alors	[ζ son]	[-ζ son]	
		[η cont]	[-η cont]	
		[θ nas]	[-θ nas]	+
		[ι voix]	[-ι voix]	

La CSM_3 se lit comme suit : si les segments en position 2,3 ne sont pas identiques, ils peuvent être homorganiques non identiques, à condition qu'ils diffèrent l'un de l'autre, soit par le son, en continuité, en nasalité ou en voisement.

Quant à la position 1,3, nous pouvons avoir la condition négative suivante :

CSM_4 :	~	+	[+ cont]	[-syll]	[+ cont]	+
		+	[+ voix]	[-syll]	[+ voix]	+

La CSM_4 négative exclut toute combinaison de segments identiques qui sont [+ cont, + voix], c'est-à-dire : /ðð, δδ, zz, γγ, ʕʕ, et ww, jj /.

D'autre part, en ce qui concerne les segments homorganiques non identiques en position 1,3, notre condition peut s'écrire de la façon suivante :

		[α ant]	[-syll]	[α ant]	
		[β cor]	[-syll]	[β cor]	
CSM_5 :	Si +	[γ haut]	[-syll]	[γ haut]	+
		[δ bas]	[-syll]	[δ bas]	
		[ϵ arr]	[-syll]	[ϵ arr]	
		[ζ son]	[-syll]	[-ζ son]	
		[η cont]	[-syll]	[-η cont]	
	Alors +	[θ voix]	[-syll]	[-θ voix]	+
		[ι C. P.]	[-syll]	[-ι C.P.]	
		[χ nas]	[-syll]	[-χ nas]	

La CSM_5 n'accepte en position 1,3, que les séquences de deux segments homorganiques non identiques, dont l'un des segments diffère de l'autre soit par le son, en continuité, en voisement, en constriction pharyngale ou en nasalité.

Ces cinq conditions qui viennent d'être présentées rendent compte de façon plus précise et plus générale de toutes les contraintes de structure morphématique de l'arabe standard. Elles permettent de distinguer entre matrices admissibles et matrices non admissibles (mots possibles et impossibles) d'une manière qui paraît naturelle.

En effet, nos conditions excluent certaines configurations non attestées qui sont incompatibles avec elles. Et comme il reste encore un nombre de configurations non attestées qui sont compatibles avec l'ensemble des conditions profondes de l'arabe standard, ce sont les *lacunes accidentelles* qui constituent les matrices admissibles et non réalisées. Ainsi, les conditions de structure morphématique nous permettent de distinguer les configurations admissibles des non admissibles, sur la base d'une extension de la méthode d'évaluation du lexique de l'arabe standard.

Les dictionnaires arabes

Pour recenser le plus grand nombre de racines attestées de l'arabe standard, nous nous sommes basés sur les trois dictionnaires arabes les plus célèbres à savoir : le dictionnaire *'Assihâh* d'El Gawhari[1], *Lisân 'Al 'Arab* d'Ibn Mandhûr[2], et *Tâg 'Al 'Arûs* d'Ezzubeidî[3].

Ces trois grands dictionnaires présentent les caractéristiques suivantes :

Dictionnaire	Année de publication (Hégire)	Racines trilitères	Racines quadrilitères	Racines quinquilitères	Total
'Assihâh	360 H.	4814	766	38	5618
Lisân Al Arab	681 H.	6538	2548	187	9273
Tâg Al 'Arûs	1200 H.	7597	4081	300	11978

Principes informatiques

La méthode de spécification SADT

Notre méthode computationnelle SADT utilise la notion de diagramme, qui est un ensemble de boîtes et de flèches étiquetées définies de la façon suivante :

1. 'Abu Nasr 'Ismâîl Bin Hammâd 'Al Fârâbî 'Al Gawharî (1984) : *'As-sihâh Tâg 'al-lughah wa Sihâh 'Al Arabiyya*, Tahqîq 'Ahmad Abdel Ghafûr, 3e éd., Beyrouth, Dar Al Ilm lilmalâyîn.
2. 'Abu 'Al Fadl Gamâl 'Ad-dîn Muhammad 'Ibn Makram 'Ibn Mandhûr (1956) : *Lisân 'Al Arab*, Beyrouth, Dar lisân 'Al Arab.
3. Muhammad Murtadâ Az-Zubeidî (1965) : *Tâg Al 'Arûs: lisarhi 'Al qâmûs Âl Muhît li Magd 'Âd-dîn 'Al Fayrûzabâdî*, 'Al Matbaah 'Al Khayriyya li Misr.

CONTRÔLE

↓

DONNÉES ENTRÉES → Nom de boîte → DONNÉES SORTIES

↑

MÉCANISME

Les données d'entrées sont celles qui vont subir le traitement destiné à récupérer les données de sortie, selon les données de contrôle. Ces dernières définissent la façon dont se déroule l'activité, tandis que les flèches de mécanisme indiquent quelles ressources (humaines ou matérielles) seront utilisées par la boîte.

La méthode SADT sépare les données des activités, ce qui implique deux types de diagrammes.

– Le DATAGRAMME représente les données et les activités en mettant l'accent sur les données. Une boîte de ce diagramme a cette forme :

ACTIVITÉ

CONTRÔLE

↓

ACTIVITÉ D'ENTRÉE → Objet → ACTIVITÉ DE SORTIE

↑

SUPPORT

– L'ACTIGRAMME est basé sur le même principe : il met l'accent sur les activités. Une boîte de ce diagamme a cette forme :

CONTRÔLE

↓

DONNÉES D'ENTRÉE → Activité → DONNÉES DE SORTIE

↑

SUPPORT

Un diagramme SADT est présenté sous forme de niveaux ; il peut y avoir entre 3 et 6 boîtes (dans chaque niveau).

La méthode de conception fonctionnelle descendante

C'est une méthode de conception dirigée par les traitements. Au cours de cette étape on définit l'architecture générale du système (du point de vue informatique).

Cette méthode comprend un **diagramme de flots des données** permettant de décrire la conception générale. Il montre comment les données sont transformées lorsqu'elles passent d'un composant du système à un autre.

Un tel diagramme est composé :
- d'arcs annotés qui représentent les flots de données en entrée et en sortie des centres de transformation ;
- de cercles annotés représentant les centres de transformation ;
- des opérateurs « * » : et « + » : ou.

Le **diagramme de structure** est une représentation hiérarchisée obtenue à partir du **diagramme de flot des données**. Il montre la relation structurelle des composants du système. Il établit la façon dont les éléments de transformation peuvent être réalisés par une hiérarchie d'unités de logiciel (module, structure).

Partant de la notion selon laquelle tout système nous permet de définir des sous-systèmes, on aboutit à l'architecture du nouveau système. Cette architecture se définit comme la **décomposition descendante d'un système**.

Cette architecture est obtenue à partir de deux décompositions :

– Une décomposition fonctionnelle qui revient à décomposer le système en sous-systèmes en essayant de regrouper dans un même ensemble des fonctions de manière cohérente.

– Une décomposition statique qui permet de décrire chaque sous-système en décrivant les différents modules de celui-ci. Elle aboutit à un sous-graphe et définit aussi l'interaction du sous-système avec les autres. Chaque module ainsi obtenu par cette décomposition peut être décomposé à son tour en unités élémentaires.

La banque des morphèmes-racines

Présentation des dictionnaires

Notre banque des morphèmes-racines de l'arabe standard comprend les dictionnaires suivants :

1. un *Dictionnaire théorique* qui contient toutes les racines trilitères théoriquement possibles, qui sont au nombre de 21 952 racines soit 28^3 (28 étant le nombre de consonnes de l'arabe standard) ;
2. un *Dictionnaire des racines admissibles,* c'est-à-dire des racines qui n'en-

freignent aucune des conditions de structure morphématique de l'arabe standard présentées plus haut ;

3. un *Dictionnaire des racines attestées,* c'est-à-dire des racines utilisées dans la langue et qui sont tirées des tableaux de répartition construits à partir des grands dictionnaires arabes ;

4. un *Dictionnaire des racines admissibles mais non attestées,* c'est-à-dire des racines qui peuvent former des mots nouveaux dans des processus de néologie lexicale ;

5. un Dictionnaire appelé : *'Al 'Ichtiqâq 'Al 'Akbar* (grande dérivation) qui, selon la terminologie du grand linguiste arabe Ibn Ginnî, consiste à générer d'une racine donnée, toutes les racines attestées (six ou moins), en permutant les trois consonnes radicales dans les trois positions de la racine trilitère. Ce dictionnaire s'avère d'une grande utilité pour vérifier l'hypothèse qui prédit une certaine parenté sémantique forte ou faible entre les six racines issues d'une racine donnée par le processus appelé : la *grande dérivation.*

Spécification

Nous présentons, dans une première étape, les *classes de données* rencontrées lors de la génération des différents dictionnaires de notre banque des racines arabes, ensuite nous présentons la *liste des activités des données* et enfin la *liste des activités.*

Les classes de données rencontrées

- une racine trilitère ;
- les 28 consonnes arabes (LC) ;
- les différents dictionnaires :
  - Dict.1 : Dictionnaire théorique ;
  - Dict.2 : Dictionnaire des racines admissibles ;
  - Dict.3 : Dictionnaire des racines admissibles et attestées ;
  - Dict.4 : Dictionnaire des racines admissibles mais non attestées ;
  - Dict.5 : Dictionnaire de la grande dérivation (GD).
- une matrice phonologique (MP) ;
- les tableaux de répartition (TR) des racines trilitères ;
- les conditions de structure morphématique (CSM).

Liste des activités des données

Nous énumérons la liste des données avec, pour chacune, les différentes activités à subir.

- **Les consonnes**
  activités : les concaténer avec d'autres.

- **Racine trilitère :**
  activités : construire la racine ;
  tester la racine en utilisant les (CSM) ;

tester la racine en utilisant les (TR) ;
classer la racine dans un dictionnaire.

– **Dictionnaire :**
activités : classer une racine dans un dictionnaire ;
consulter un dictionnaire.

– **Matrice phonologique (MP) :**
activités : consulter la matrice.

– **Tableau de répartition des racines trilitères (TR) :**
activités : saisir un tableau ;
consulter un tableau.

– **Les conditions de structure morphématique (CSM) :**
activités : appliquer une CSM à une racine.

– **Le Dictionnaire : grande dérivation (GD) :**
activités : classer une racine ;
consulter le dictionnaire.

Listes des activités

Nous avons pu établir une dizaine d'activités qui sont :

1. construire une racine trilitère ;
2. saisir la matrice phonologique (MP) ;
3. consulter la matrice phonologique (MP) ;
4. tester une racine par les conditions de structure morphématique (CSM) ;
5. saisir les tableaux de répartition (TR) ;
6. consulter un tableau de répartition (TR) ;
7. tester une racine par un tableau de répartition ;
8. classer une racine dans un dictionnaire ;
9. construire le dictionnaire de la grande dérivation ;
10. consulter un dictionnaire.

Enfin, nous présentons ci-dessous le DATAGRAMME et l'ACTIGRAMME qui regroupent les données et les activités rencontrées dans ce module, ainsi que la spécification sous forme de graphique.

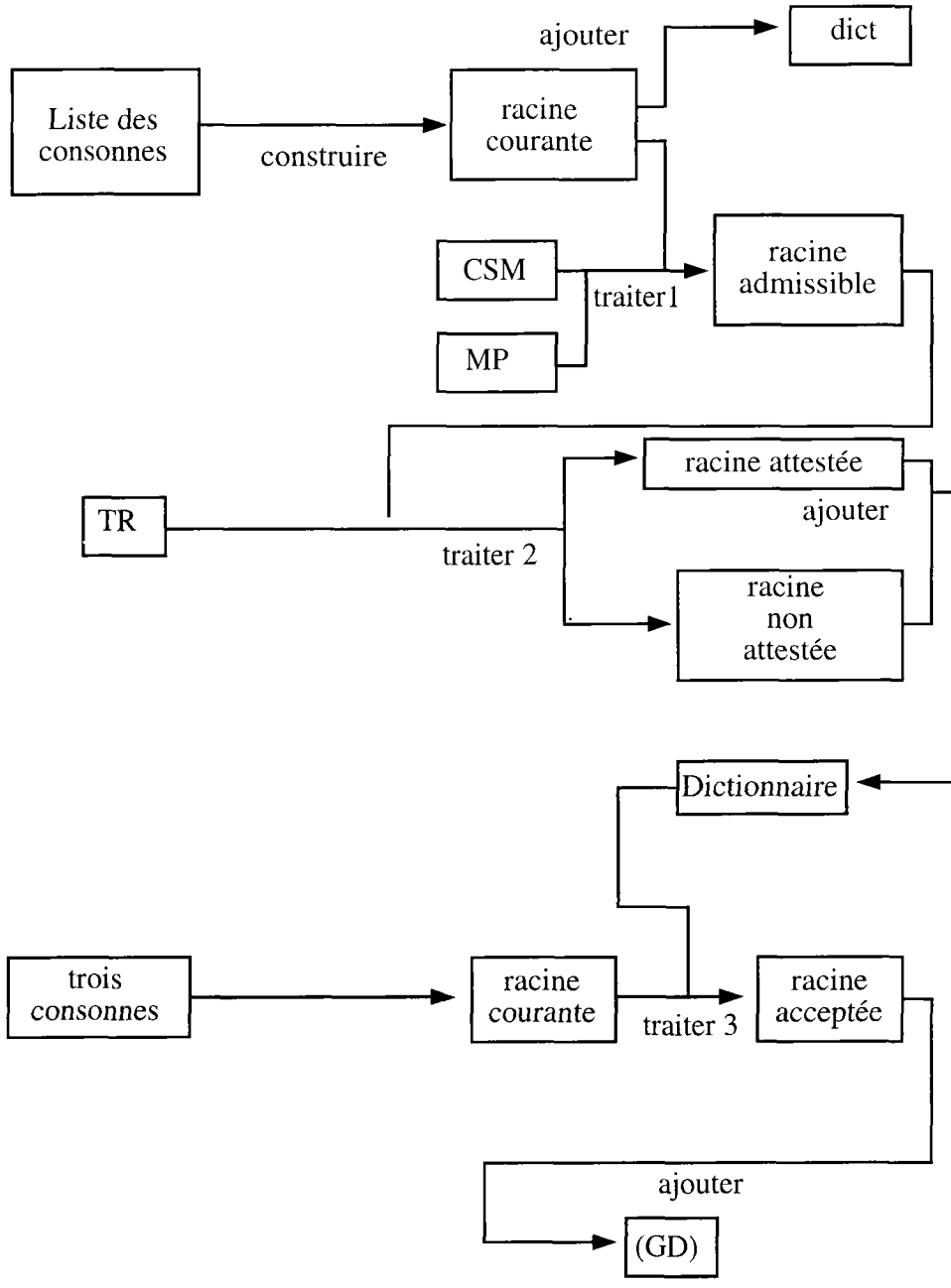
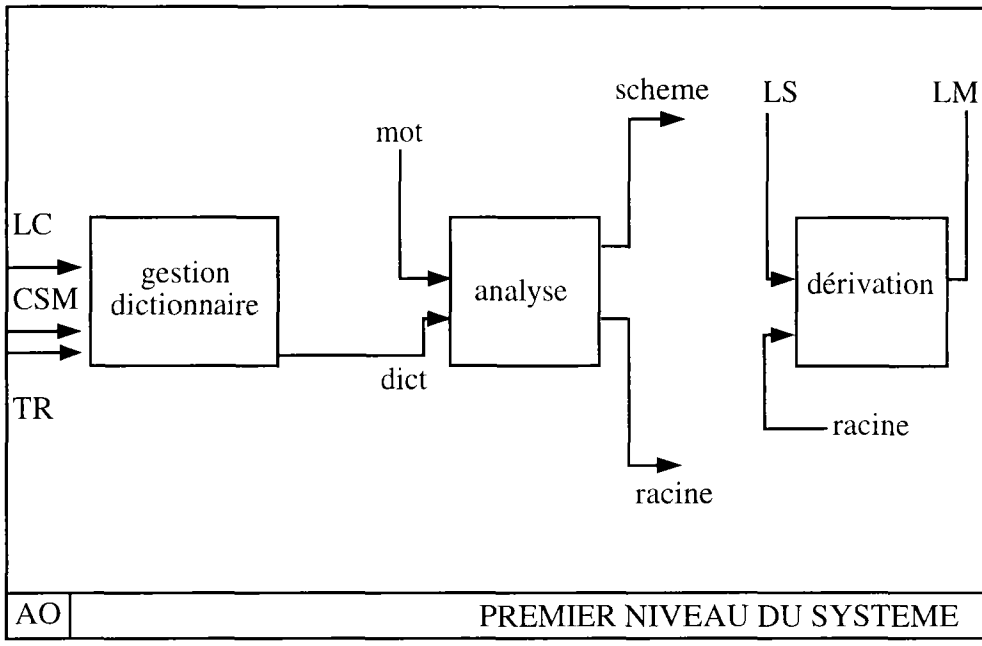
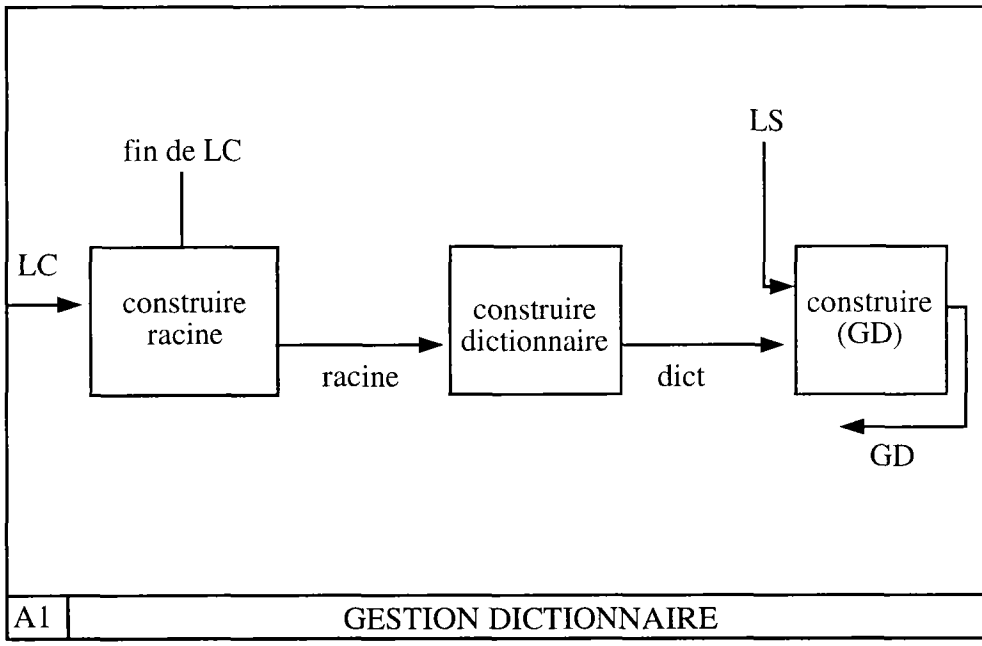
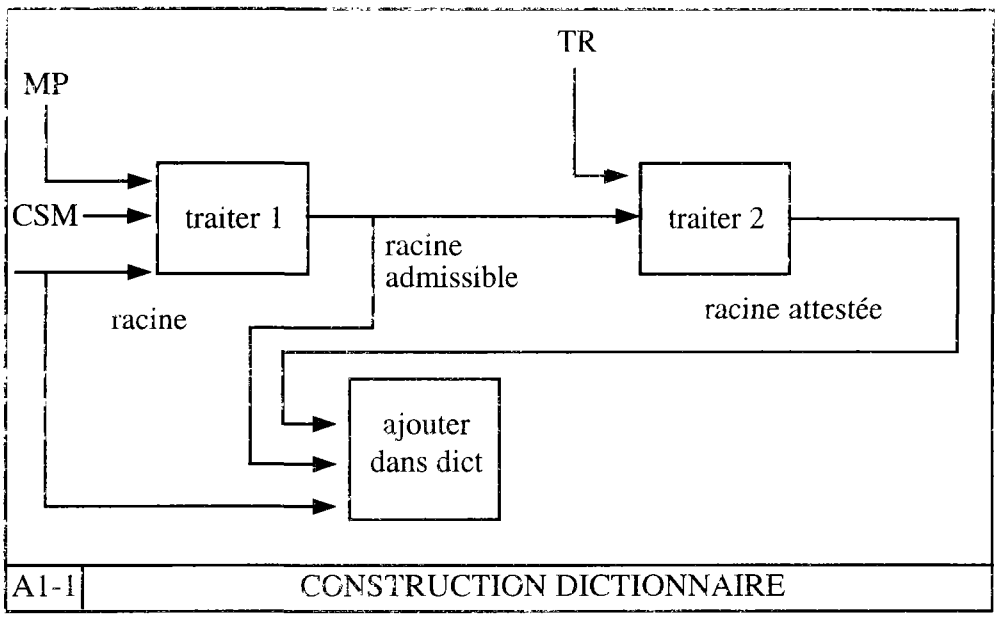
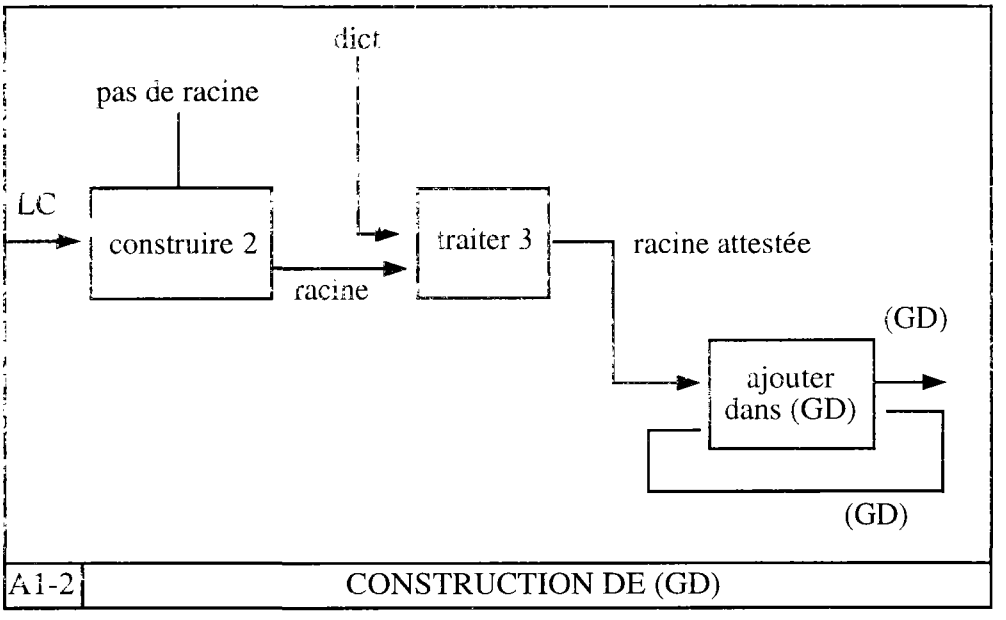


Diagramme des Données relatives au module DICTIONNAIRES


scheme
LS
LM
mot
LC
gestion dictionnaire
analyse
dérivation
CSM
dict
TR
racine
racine
AO
PREMIER NIVEAU DU SYSTEME


LS
fin de LC
LC
construire racine
construire dictionnaire
construire (GD)
racine
dict
GD
A1
GESTION DICTIONNAIRE





Étape de conception

Partant des résultats de l'étape de spécification, nous présentons ici les différents traitements nécessaires pour cette étape de conception fonctionnelle descendante.

Traitements utilisés

Nous donnons ci-dessous les différentes activités nécessaires pour la réalisation de ce module, ainsi que leurs entrées, sorties et principes.

- Construire racine
  - entrée : liste des consonnes ;
  - sortie : une racine trilitère.
- Appliquer les CSM à une racine
  - entrée : racine ;
    (MP) ;
    (CSM) ;
  - sortie : racine admissible.
  - principe : appliquer les (CSM) à la racine pour la tester.

- Tester si la racine est attestée
  - entrée : racine admissible ;
    tableau de répartition.
  - sortie : racine attestée ou non.
  - principe : en utilisant le tableau de répartition correspondant à la première consonne de la racine pour voir si cette dernière est attestée ou non.
- Ajouter dans un dictionnaire
  - entrée : dictionnaire ;
    racine.
  - sortie : dictionnaire.
  - principe : ajouter la racine à la fin du dictionnaire en conservant l'ordre alphabétique.
- Construire 2 racine
  - entrée : trois consonnes distinctes.
  - sortie : racine trilitère.
  - principe : construire une racine dont les consonnes sont différentes.

- Traiter 3
  - entrée : racine trilitère ;
    dictionnaire des racines attestées.
  - sortie : racine acceptée ou non.
  - principe : tester si la racine en entrée est dans le dictionnaire des racines attestées ou non.
- Ajout dans le dictionnaire (GD)
  - entrée : racine trilitère ;
    dictionnaire (GD).
  - sortie : dictionnaire (GD).
  - principe : même principe que l'ajout dans Dict.

Diagramme de flots de données (DFD)

Nous présentons ci-dessous le diagramme de flots de données concernant la génération des dictionnaires.

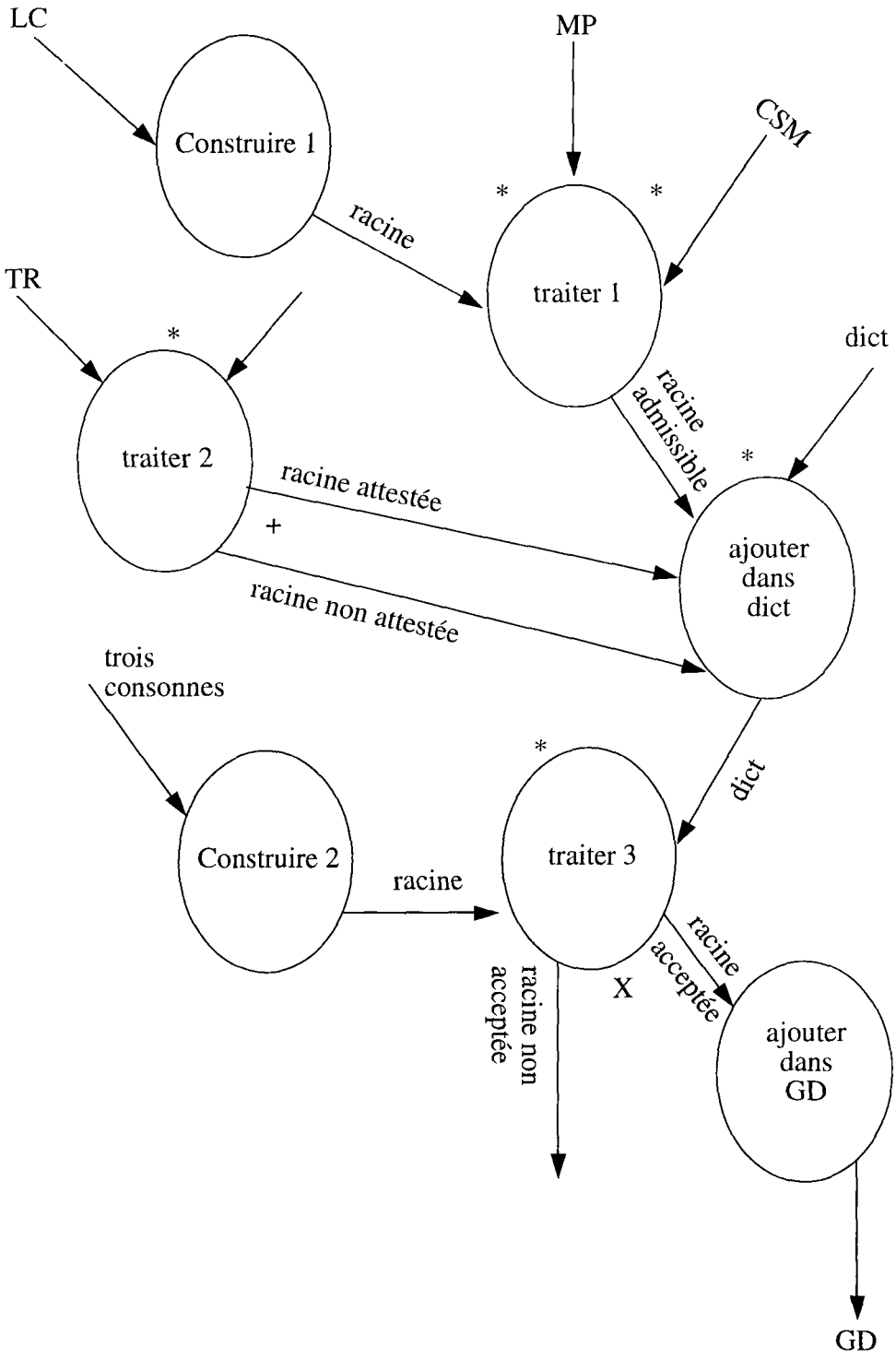
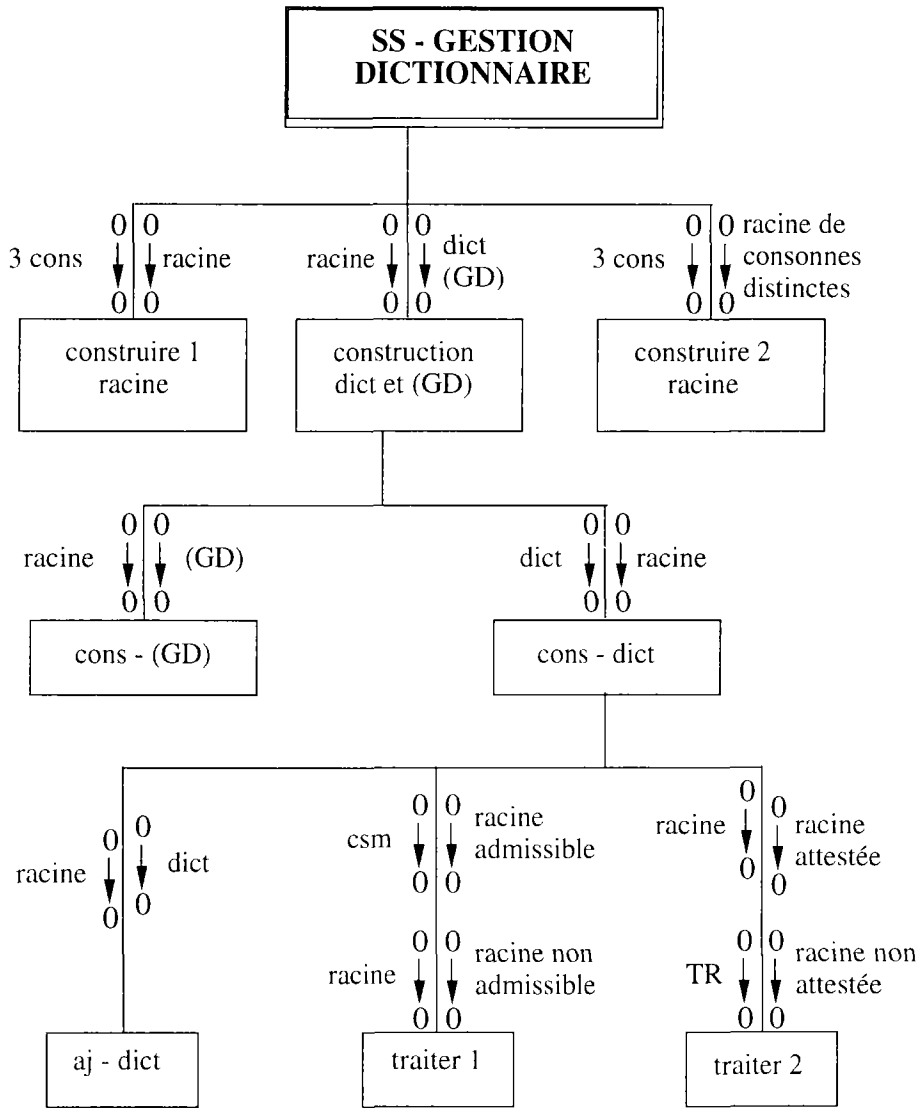

LC
Construire 1
MP
CSM
racine
*
*
traiter 1
TR
*
traiter 2
dict
racine admissible
*
racine attestée
+
racine non attestée
ajouter dans dict
trois consonnes
*
Construire 2
racine
traiter 3
dict
racine acceptée
X
racine non acceptée
ajouter dans GD
GD

Diagramme de structure

Nous présentons ci-dessous le diagramme de structure :



Remarque :

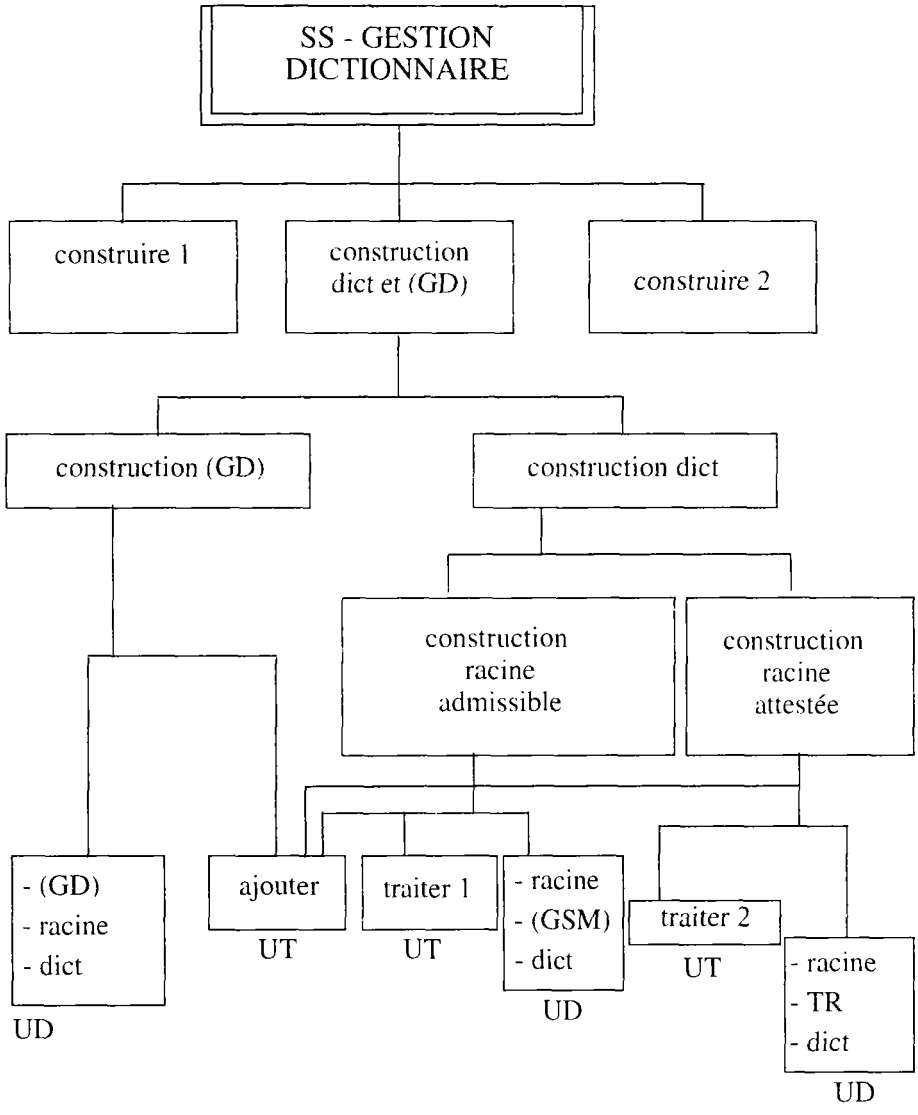
aj-dict : ajouter dans le dictionnaire.
cons-(GD) : construire (GD).
cons-dict : construire dictionnaire.
CSM : conditions de structure morphématique.
TR : tableaux de répartition des racines trilitères.

Décomposition descendante du système

Notre système peut être subdivisé en trois sous-systèmes, à savoir :

- un sous-système de génération des dictionnaires ;
- un sous-système d'analyse ;
- un sous-système de dérivation.

On aboutit ainsi à l'architecture du nouveau système qui se définit de la façon suivante :



Formalisme informatique

Nous décrirons le formalisme informatique adopté pour réaliser les objectifs fixés au départ, à savoir la génération automatique des différents dictionnaires de notre banque des morphèmes-racines de l'arabe standard.

Les structures de données utilisées

a. La matrice phonologique

Pour implémenter cette matrice, on a utilisé la structure de données : Tableau bi-dimensionnel (2, 28) : les deux lignes représentent respectivement les deux traits [cont, voix], et les 28 colonnes représentent les 28 consonnes.

Le terme général de ce tableau est C [i, j] (j = 0..27) (i = 0..1). C [i, j] = 1 si la consonne j est [cont ou voix] (selon que i = 0 ou 1). C [i, j] = 0 si la consonne j n'est pas [cont] (si i = 0) n'est pas [voix] (si i = 1).

b. Les tableaux de répartition

Pour la génération du dictionnaire attesté ont été utilisés 28 tableaux de répartition. L'étude de leur structure a généré un certain nombre de remarques, qui vont nous guider dans le choix d'une implémentation adéquate de ces tableaux, aux fins d'optimiser l'occupation en espace mémoire et temps de réponse.

Les remarques tirées sont les suivantes :

– Les tableaux de répartition sont des matrices binaires, dans lesquelles l'occurrence de la valeur 1 est nettement inférieure à celle de la valeur 0, raison pour laquelle on peut les considérer comme des matrices creuses.

– On dispose pour chaque ligne, ainsi que pour chaque colonne du nombre de valeurs égales à 1.

c. Mise en œuvre informatique proposée

Nous proposons de présenter chaque tableau de répartition par un **fichier-texte** et en utilisant une structure bien précise. Nous donnons comme exemple de fichier le tableau de répartition des racines trilitères à première consonne radicale /ʔ/ :

	ʔ	ʕ	b	t	θ	ʒ	ħ	x	d	ð	r	z	s	ʃ	S	D	T	δ	ʕ	γ	f	q	k	l	m	n	h	w	j
2	ʔ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	b	0	1	1	1	0	0	0	1	0	1	1	1	0	0	1	1	0	0	1	0	1	0	1	0	1	1	1		
	t	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1		
	θ	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1			
	ʒ	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0		
	ħ	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		
	x	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
	d	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1		
	ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1		
	r	0	1	0	1	1	0	1	0	0	1	1	1	1	0	1	1	0	0	0	1	1	1	0	1	1	0	1		
	z	0	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1		
	s	0	1	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1	1	1	0	1		
	ʃ	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1		
	S	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1		
	D	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1		
	T	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0		
	δ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
	ʕ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
	γ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
	f	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0			
	q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0			
	k	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0			
	l	0	1	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	1	1	1	1	0	1	1			
	m	0	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	1	1	1			
	n	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1			
	w	1	1	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	1	1	0	1	1	1	1			
	h	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0		
	j	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0	1	1	1	1	1	1		

FICHIER N0

00
16 : ي : و : ه : ن : ل : ق : غ : ط : ض : س : ز : ر : د : ث : ت : ب
08 : ي : و : ه : ن : ل : م : ل : ت : ب
07 : ي : و : م : ل : ف : ر : ث
08 : ن : م : ل : ص : ر : د : ج : أ
03 : ن : د : ح
04 : ي : و : ر : ن
07 : ي : و : م : ل : ر : د : ب
03 : ي : و : ن
17 : ي : و : ن : م : ك : ق : ف : ط : ض : ش : س : ز : ر : خ : ج : ث : ب
12 : ي : و : م : ل : ق : ف : ز : ر : د : ح : ج : ب
12 : ي : و : ن : م : ل : ك : ك : ف : س : ر : د : ت : ب
06 : ي : و : ف : ش : ر : ب
07 : ي : و : ل : ف : ص : ر : د
05 : ي : و : م : ض : خ
04 : م : ل : ط : ر
00
00
00
08 : ن : ل : ك : ق : ف : ر : د : خ
04 : ه : ن : ط : ر
06 : م : ل : ك : ف : ر : د
14 : ي : و : ه : ل : ك : ق : ف : ض : س : خ : ح : ت : ب
12 : ي : و : ه : ن : م : ل : ع : س : ر : د : ج : ت
11 : ي : و : ه : ن : ك : ق : ف : س : ث : ت : ب
14 : ي : و : ه : ن : م : ل : ق : ف : س : ز : ر : د : ب : أ
05 : ن : ل : ق : ر : ب
11 : ي : و : ه : ن : م : ل : ك : ض : س : ن : خ

Conclusion

Notre banque permet de réaliser les fonctions suivantes :

- la génération automatique de cinq dictionnaires des racines trilitères de l'arabe standard (dictionnaire théorique, dictionnaire des racines attestées, dictionnaire des racines admissibles, dictionnaire des racines non admissibles, dictionnaire de la grande dérivation) ;
- la consultation de ces dictionnaires, dans le but de tester une racine trilitère ;
- l'affichage d'une liste des racines d'un dictionnaire spécifié ;
- l'affichage des racines attestées dont les consonnes radicales sont permutables en position 1,2,3 : *la grande dérivation.*

Références

'AL GAWHARÎ, 'Abu Nasr 'Ismâîl Bin Hammâd 'Al Fârâbî (1984) : 'As-sihâh: Tâg 'al-lughah wa Sihâh 'Al Arabiyya, Tahqîq 'Ahmad Abdel Ghafûr, 3e éd., Beyrouth, Dar Al Ilm lilmalâyîn.

AZ-ZUBEIDÎ, Muhammad Murtadâ (1965) : *Tâg Al 'Arûs: lisarhi 'Al qâmûs Âl Muhît li Magd 'Ad-dîn 'Al Fayrûzabádî*, 'Al Matbaah 'Al Khayriyya li Misr.

CHOMSKY, Noam (1968) : *The Sound Pattern of English*, New York, Harper and Row.

HABAILI, Hussein (1976) : *Contraintes de structures morphématiques en arabe,* mémoire de maîtrise ès arts en linguistique, Montréal, Université de Montréal.

HABAILI, Hussein (1990) : *Phonologie générative et morphologie flexionnelle et dérivationnelle de l'arabe.* Thèse de doctorat d'État, Paris, Université de la Sorbonne nouvelle.

'IBN MANDHUR 'abu 'Al Fadl Gamâl 'Ad-dîn Muhammad 'Ibn Makram (1956) : *Lisân 'Al Arab*, Beyrouth, Dar lisân 'Al Arab.

STANDLEY, Robert (1976) : « Redundancy Rules in Phonology », *Language*, vol. 43, pp. 393-436.

32

L'enseignement de la traduction franco-malgache assisté par ordinateur ou appuyé par la traductique

Roger-Bruno RABENILAINA

Université d'Antananarivo, Antananarivo, Madagascar

Introduction

La TAO n'est pas encore appliquée ni applicable, *stricto sensu,* à la traduction franco-malgache. La raison en est simple : une description détaillée de la syntaxe du malgache n'est pas encore disponible, alors que notre coopération avec les équipes du LADL de Paris 7, et du LLI de Paris 13, nous fait largement profiter d'une bonne couverture de celle du français. Seuls le verbe et l'adverbe ont bénéficié jusqu'ici, en malgache, de recherches suffisamment poussées (voir Rabenilaina 1985, 1991, 1993 et Raharinirina-Rabaovololona 1991). Des travaux de type doctoral sont actuellement en cours sur l'adjectif, le nom prédicatif, le verbe composé, le verbe neutre, le verbe réciproque, l'enseignement du malgache langue maternelle et le déterminant au niveau de l'objet direct du verbe. Un essai de codification des règles morphologiques et prosodiques, qui président à la dérivation verbale par préfixation et suffixation, a été entrepris en 1990 à partir des manuels de grammaire traditionnelle et des thèses d'État de morphosyntaxe soutenues jusqu'alors ; mais cet essai attend encore l'intervention d'un programmeur.

Nous n'avons pas attendu, toutefois, l'informatisation complète de la grammaire malgache pour faire appel à l'ordinateur dans l'enseignement de la traduction franco-malgache dont nous sommes chargé à l'Université d'Antsiranana depuis 1986. De fait, par souci d'économie et puisque nous avions à traduire et à faire traduire à nos étudiants un grand nombre de textes variés, nous avons commencé par nous constituer deux types de banques terminologiques, dont l'une est à entrées françaises et l'autre à entrées malgaches, les mots étant rangés par ordre alphabétique de part et d'autre. Au

fur et à mesure que nous traduisons des textes littéraires (qui peuvent être oraux en malgache) ou journalistiques, que nous entreprenons des recherches dans ce sens et que nous adoptons enfin des équivalents (français, en version, et malgaches, en thème) en langue d'arrivée, nous notons ces derniers avec toutes leurs particularités syntaxiques et sémantiques. Nous créons une fiche pour la version et une autre pour le thème et nous entrons les termes dans la banque terminologique correspondante ainsi constituée. Lorsque nous nous heurtons de nouveau au même groupe de mots dans un texte ultérieur, nous pouvons consulter ces fiches informatisées en évitant ainsi la même série de recherches. C'est dans ce sens, et dans ce sens seulement, que doit être compris le titre que nous avons donné à cette communication.

Nos explications en classe consistent ainsi, au cours de l'exercice de traduction, à faire prendre surtout conscience à nos étudiants des distances grammaticales et lexicales qui séparent le français et le malgache. Nous ne pourrons pas traiter ici de toutes les différences qui opposent les deux langues. Nous concentrerons notre exposé sur ce qu'il est convenu d'appeler « modalités grammaticales », à savoir : le nombre, le genre, la personne et la voix, en insistant particulièrement sur cette dernière. C'est au niveau du traitement de ces modalités que se situent, à notre avis, les principales difficultés auxquelles nous sommes confronté dans l'enseignement de la traduction franco-malgache et, sans doute aussi et de façon plus cruciale, en TA et TAO. Nous comptons donc sur vos suggestions sur ce dernier point.

Les modalités de nombre, de genre et de personne

Il n'y a ni nombre, ni genre, ni personne en malgache, contrairement au français. Il faudra formaliser et coder ces différences en traductique de façon que la machine soit capable de les traiter automatiquement en passant d'une langue à l'autre.

Le nombre

Toutes les catégories grammaticales ou parties du discours sont invariables, hormis le pronom personnel (*Pro*) et le pronom ou adjectif démonstratif (*Dém*). Il s'agit du verbe (*V*), de l'adjectif (*Adj*), du nom (*N*) et de l'article (*Dét*), pareillement à l'adverbe (*Adv*), à la conjonction (*Conj*), à la préposition (*Prép*), etc. Les exemples suivants, bâtis sur le français, sont caractéristiques :

V =: *lire*	= *mamaky* :
je lis	= mamaky aho
tu lis	= mamaky ianao
(il, elle) lit	= mamaky izy
nous lis*ons*	= mamaky (isika *incl*, izahay *excl*)
vous lis*ez*	= mamaky ianareo
(ils, elles) lis*ent*	= mamaky izy ireo
Adj =: grands, grand	= lehibe
N =: arbres, arbre	= hazo
Dét =: les étudiants, l'étudiant	= ny mpianatra

En ce qui concerne les exceptions, l'exemple du verbe *lire* = *mamaky* conjugué ci-dessus montre que le pronom personnel présente aussi différentes formes en malgache selon qu'il est au singulier ou au pluriel. On remarquera qu'à la première personne du pluriel il existe, en malgache, une forme inclusive = *incl* (l'auditoire est inclus dans le procès) et une forme exclusive = *excl* (l'auditoire est exclus du procès). Quant au pronom ou adjectif démonstratif, ses formes au pluriel résultent de l'infixation de *-re-* aux formes du singulier ; exemples :

Dém =:	ce, cette	= io, ito (singulier)
	ces	= i*re*o, i*re*to (pluriel)
	(celui, celle)-là	= iny (singulier)
	(ceux, celles)-là	= i*re*ny (pluriel)

Le genre

À l'opposé du français, l'adjectif, le nom, l'article et le démonstratif n'ont pas de genre en malgache. Ils ne se mettent donc ni au masculin ni au féminin. Ils sont, pourrait-on dire, de genre neutre. On a les exemples suivants déjà cités :

Adj	=: grand, grande	= lehibe
N	=: étudiant, étudiante	= mpianatra
Dét	=: le, la	= ny
Dém	=: ce, cette	= io

La personne

Le verbe malgache n'a pas de conjugaison, au sens où les personnes s'expriment chacune par une terminaison particulière. C'est ce qu'indiquent les exemples construits sur le verbe *lire* = *mamaky*, où les pronoms personnels se distinguent en 1re, 2e et 3e personne du singulier et du pluriel sans que le verbe change de désinence.

En résumé, les expressions morphologiques des notions de genre, de nombre et de personne sont inexistantes en malgache. Cette particularité pose des problèmes de représentation dans une traduction franco-malgache automatique ou assistée. Comment résoudre ces problèmes ? Nous laissons la question sans réponse pour passer immédiatement aux difficultés plus cruciales que présente la modalité de « voix ».

La modalité de voix

Le verbe malgache présente trois voix : l'active, la passive et la relative, la pronominale, qui n'a pas de forme propre, étant intégrée à l'active.

Comme ces trois voix comportent une facette formelle et une facette fonctionnelle, nous désignons la première facette par « voix » et la seconde par « diathèse ».

La voix est, dans ce sens, une catégorie morphologique et la diathèse une catégorie sémantique. Nous examinerons cette dernière plus loin pour ne nous occuper

ici que de la première. Signalons qu'en malgache la voix active est moins employée que les voix passive et relative, alors qu'elle est de loin préférée à la passive en français.

La voix active

Le verbe actif comporte un préfixe complexe désigné par *m-x-*, où *m-* exprime en même temps le présent par opposition à *n-* le passé et à *h-* le futur qui alternent avec lui, et où *x-* varie avec les radicaux :

m-x- =: m-zéro	: m-/n-/h-ita	= traverser présent/passé/futur
m-x- =: m-a-	: m-/n-/h-a-tory	= dormir présent/passé/futur
m-x- =: m-am-	: m-/n-/h-am-aky	= lire présent/passé/futur
m-x- =: m-an-	: m-/n-/h-an-deha	= aller présent/passé/futur
m-x- =: m-i-	: m-/n-/h-i-jery	= regarder présent/passé/futur

Des règles précises gouvernent la combinaison du préfixe *m-x-* avec le radical, selon le type d'initiale vocalique ou consonantique qui commence ce dernier (voir Rabenilaina 1991). Il en est de même des changements morphologiques et prosodiques qui accompagnent la suffixation de *-ana* et de *-ina* dans la formation du passif et du relatif.

La voix passive

Le verbe à la forme passive est caractérisé par un affixe simple, considéré soit à l'aspect perfectif (*voa-* et *-in-*), soit à l'aspect imperfectif (*a-, -ana* et *-ina*). Il semble, cependant, que le préfixe *a-* et les suffixes *-ana* et *-ina* dénotent aussi l'aspect perfectif, toutes les fois que le radical verbal n'est pas compatible avec les affixes perfectifs *voa-* et *-in-*. Nous l'indiquerons par une étoile (*) dans les exemples qui suivent, quoique des recherches lexico-syntaxiques plus poussées soient encore nécessaires sur ce point :

- le passif à préfixe *a-* :
  toro (action de montrer)
  = *a-toro* (qu'on montre)
  = *voa-toro, t-in-oro* (qu'on a montré)
  leha (marche)
  = *a-leha* (où l'on marche)
  = **voa-leha, *l-in-eha* (où l'on a marché)

- le passif à suffixe *-ana*
  fafa (nettoyage)
  = *faf-ana* (qu'on nettoie)
  = *voa-fafa, f-in-afa* (qu'on a nettoyé)
  tolotra (offre)
  = *tolor-ana* (à qui l'on offre)
  = **voa-tolotra, t-in-olotra* (à qui l'on a offert)

mais on a :

a-tolotra (qu'on donne)
= *voa-tolotra, t-in-olotra* (qu'on a donné)

- le passif à suffixe *-ina* :
lainga (mensonge)
= *lainga-ina* (à qui l'on ment)
= *voa-lainga, l-in-ainga* (à qui l'on a menti)
vidy (achat)
= *vid-ina* (qu'on achète)
= *voa-vidy, v-in-idy* (qu'on a acheté)

mais on a :

vidy (vente)
= *a-vidy* + *am-idy* (qu'on vend)
= **voa-vidy,*v-in-idy* (qu'on a vendu)

On remarque, à travers ces exemples, qu'un même mot, qui recouvre deux radicaux verbaux, peut convoyer deux aspects différents. Tel est le cas de *tolotra* (offre, don) et de *vidy* (achat, vente). C'est ce qui nous fait dire qu'un inventaire exhaustif de tous les emplois d'un radical verbal s'impose dans le cadre du lexique-grammaire du malgache, dont la construction s'impose avant d'entreprendre la TA et la TAO franco-malgaches.

La voix relative

La forme du verbe tient ici à la fois de l'actif et du passif. Elle est ainsi caractérisée par un affixe discontinu, appelé circonfixe, *x-...-ana*, dont l'élément préfixal appartient à l'actif et l'élément suffixal au passif. Nous avons les exemples suivants :

m-am-aky =: *mamaky* (lire) = *am-aki-ana* =: *amakiana* (où, avec quoi, à qui...lire) /*x-* =: *am-*
m-i-jery =: *mijery* (regarder) = *i-jeri-ana* =: *ijerena* (où, avec quoi, pour qui... regarder) / *x-* =: *i-*
m-a-tory =: *matory* (dormir) = *a-tori-ana* =: *atoriana* (où, avec quoi, pour qui... dormir) / *x-* =: *a-*
m-zéro-ita =: *mita* (traverser) = *zéro-ita-ana* =: *itana* (où, avec quoi...traverser) /*x-* =: zéro

La traduction que nous proposons dans ces exemples n'est qu'approximative, car le sens du verbe est tributaire de la structure de la phrase et vice versa.

En effet, le recours aux différentes voix qu'on vient de parcourir indique seulement que le locuteur veut « pointer » :

- sur l'agent, dans l'actif, en le focalisant en sujet « grammatical » ;
- sur un complément, dans le passif et le relatif, en le focalisant en sujet « grammatical ».

La détermination du contenu sémantique ou fonctionnel de la relation verbe-complément (*V-Comp*) est donc cruciale pour la traduction. Elle se fait habituellement au niveau de l'actif, où les différentes prépositions qui introduisent les compléments sont apparentes. D'où le passage de la modalité de voix à celle de diathèse, que nous avons adoptée depuis 1985.

La modalité de diathèse

Il faut distinguer, d'emblée, deux types de diathèses en malgache. Le premier type comprend les diathèses « de base » et le second les diathèses « dérivées ». Les phrases sont à l'actif dans le premier cas : le sujet « logique » est sujet grammatical, c'est-à-dire que l'agent est focalisé. Les phrases sont, par contre, au passif et/ou au relatif, dans le second cas : le complément est sujet grammatical, c'est-à-dire que l'« objet » (direct ou indirect) ou le « circonstanciel » est focalisé. Nous passerons en revue les deux types de diathèses en question.

Les diathèses de base

Nous avons dénombré quatre diathèses à l'actif : la transitive, l'intransitive, la factitive et la réciproque. Les deux premières concernent généralement des phrases simples à au moins deux arguments, alors que les deux dernières ne se rencontrent qu'avec des phrases complexes à trois arguments au minimum.

Nous rappelons (voir Rabenilaina 1985, 1991) qu'un verbe malgache se définit par sa combinatoire morphologique (présence d'un préfixe *m-x-* transformable en un circonfixe *x-...-ana*) et sa combinatoire distributionnelle (présence d'un groupe *nominal* ou *prépositionnel* non antéposable à sa droite et dont la tête désigne un humain et/ou un non humain selon les items). En termes fonctionnels, un élément verbal prédicat, à l'actif, comporte donc ou peut toujours comporter un complément d'objet direct ou indirect en malgache, contrairement au verbe français, qui, selon une certaine tradition grammairienne, n'a pas de complément à l'intransitif. Il est ainsi possible de distinguer plusieurs types de diathèse intransitive en malgache, alors qu'il n'existe qu'un seul type de diathèse transitive, quand bien même un verbe transitif peut et doit souvent s'accompagner d'un complément d'objet indirect pour lever toute ambiguïté.

La diathèse transitive

Une phrase (à la diathèse) transitive a comme premier complément un groupe nominal construit directement sur le verbe. Autrement dit, $N_1 =: N$ est un objet direct « patient » ou « passif ». Exemple :

Mamangy V	*ny reniny* N_1	*i Soa* N_0
Visite *V*	la mère d'elle N_1	(la fille) Soa N_0
Soa visite sa mère		

La diathèse intransitive

Une phrase (à la diathèse) intransitive a son premier complément en construction indirecte avec le verbe. On dit alors que N_1 *=: Prép N* est un objet indirect « non patient » ou « non passif ». Tout objet indirect pouvant comporter la préposition « passe-partout » *Prép =: am =: amy + amina,* la valeur ou la portée de sa relation avec le verbe est caractérisée par la ou les variantes de *Prép =: am.* On a ainsi inventorié les types d'objet indirect suivants. (Nous signalons, pour mémoire, que de tels objets indirects se présentent aussi dans une phrase transitive, en position de N_2 *=: Prép N:* des exemples en seront proposés plus loin).

Objet instrumental. La préposition est effaçable (*Prép =: E + am*) devant un complément d'instrument. L'absence de *Dét =: ny* (le, la, les) est alors obligatoire. Exemple :

Misarotro V (*E + amin'ny*) *tsihy* N_1 *i Be*
Se-protège-de-la-pluie *V* avec (un + la) natte N_1 Be N_0
Be se-protège-de-la-pluie avec (une + la) natte

Objet destinataire humain. La préposition *Prép =: am* est obligatoire devant un complément destinataire humain. Exemple :

Mandainga V *amin'i Be* Prép N_1 *i Soa* N_0
Ment *V* à Be *Prép* N_1 Soa N_0
Soa ment à Be

Objet destinataire non humain. La préposition *Prép =: am* est commutable avec la variante *Prép =: Loc am =: (ato, eo, ao, any, eny) am* ((ici, là) (à, dans, ...)). Exemple :

Mandeha V *eny amin'ny arabe* Prép N_1 *ny fiara* N_0
Roule *V* là à la chaussée *Prép* N_1 la voiture N_0
La voiture roule sur la chaussée

Objet datif. La préposition *Prép =: am* est obligatoire devant un complément datif. Celui-ci, qui est humain, est, comme nous le verrons plus bas, distingué du destinataire humain par le fait que, focalisé en sujet grammatical, il exige que le verbe soit uniquement à la forme relative, alors qu'avec le destinataire humain le verbe peut ou doit être à la forme passive selon les items lexicaux. Exemple :

Miresaka V *amin'i Soa* Prép N_1 *i Be* N_0
Parle *V* à Soa *Prép* N_1 Be N_0
Be parle à Soa

Objet locatif. La préposition *Prép =: am* peut alterner avec les variantes *Prép =: Loc (am + an) =: (ato, eto, eo, ao, eny, any) (am + an)* ((ici, là) (à, dans, ...)). Ici aussi, comme nous y reviendrons ci-dessous, l'objet locatif s'oppose au destinataire non humain en ce que, focalisé, il ne gouverne que la voix relative, contrairement au destinataire non humain, qui accepte en même temps la voix passive. Exemple :

Monina V (*E* + (*E* + *eny*)) (*amy* + *an-*) *lohasaha* Prép N_1 *ry Be* N_0
Habitent *V* (E + là) (à + dans) vallée *Prép* N_1 les Be N_0
Les Be habitent dans (une + la) vallée

Objet de moyen. La préposition *Prép =: am*, effaçable, est équivalente à la locution prépositive *Prép =: am al =: amin'ny alalan'ny* (au moyen de, par l'intermédiaire de) devant un objet de moyen. Outre le fait qu'il accepte une telle locution, celui-ci diffère d'un objet instrumental : focalisé, il est le sujet d'un verbe relatif, alors que l'instrument peut être aussi celui d'un verbe passif. Exemple :

Mandeha V (*E* + *amin'ny* (*E* + *alalan'ny*)) *fiara* N_1 *i Be* N_0
Voyage *V* (E + par le moyen de la) voiture N_1 Be N_0
Be voyage en voiture

La diathèse factitive

Il est possible, dans le cadre d'une phrase active, de faire intervenir, de l'extérieur du procès, un second sujet, qu'on appelle, pour cette raison, agent extérieur, noté N_e. La phrase obtenue est désignée du nom de phrase factitive (causative, dans la terminologie traditionnelle). Tous les verbes actifs, transitifs ou intransitifs, sont ainsi passibles de factitivation, moyennant, en plus, l'insertion de l'infixe *-amp-* (faire, rendre) entre *m-* et *x-* ; d'où le préfixe complexe : *m-amp-x-*. L'agent extérieur, en position de sujet grammatical, est le « causateur » et l'agent véritable, en position de complément, le « causé ».

Soient les deux phrases transitive et intransitive suivantes :

Miantso V *ny ankizy* N_1 *i Soa* N_0
Appelle *V* les enfants N_1 Soa N_0
Soa appelle les enfants

Mandainga V *amin'ny ankizy* N_1 *i Soa* N_0
Ment *V* à les enfants N_1 Soa N_0
Soa ment aux enfants

Si nous infixons *-amp-* au verbe et que nous plaçons à la fin de chaque phrase l'agent extérieur N_e *=: i Be* (nom d'un garçon), le sujet N_0 *=: i Soa* (nom d'une fille) se fait précéder du « proclitique d'objet » *an'* (exigé par un nom propre) et se place après le N_1 *=: ny ankizy* (les enfants) dans la transitive et avant lui dans l'intransitive. Nous obtenons les phrases à la diathèse factitive (ou causative) suivantes :

Mampiantso V *ny ankizy* N_1 *an'i Soa* N_0 *i Be* N_e
Fait appeler *V* les enfants N_1 Soa N_0 Be N_e
Be fait appeler les enfants par Soa

Mampandainga V *an'i Soa* N_0 *amin'ny ankizy* N_1 *i Be* N_e
Fait mentir *V* Soa N_0 à les enfants N_1 Be N_e
Be fait mentir Soa aux enfants

On voit bien qu'une phrase factitive est une phrase complexe.

La diathèse réciproque

La phrase réciproque est complexe d'une autre façon. Elle suppose la coordination de deux phrases qui, toutes choses égales d'ailleurs, comportent des têtes identiques dans les groupes nominaux nucléaires de même distribution ou fonction. Le verbe n'étant pas répété, N_1 et N_0 sont connectés par la conjonction de coordination *Co =: sy* (et), sinon le N_i (soit N_1, soit N_0, au choix) le plus proche de *V* est introduit par la préposition *Prép =: am.* On a alors recours à l'infixe simple *-if-*, quand *x- =: an-, am-*, et à l'infixe complexe *-ifamp-*, quand *x- =: i-, zéro.* Nous aurons ainsi les phrases réciproques suivantes, en travaillant dans ce sens sur les phrases transitives proposées dans le paragraphe sur la diathèse factive :

Mifampiantso V *ny ankizy* N_1 *sy* Co *i Soa* N_0
S'appellent *V* les enfants N_1 et *Co* Soa N_0
Les enfants et Soa s'appellent entre eux
=
Mifampiantso V *amin'ny ankizy* Prép N_1 *i Soa* N_0
S'appellent *V* avec les enfants *Prép* N_1 Soa N_0
Soa s'appelle avec les enfants

Mifandainga V *amin'ny ankizy* Prép N_1 *i Soa* N_0
Se ment *V* avec les enfants *Prép* N_1 Soa N_0
Soa se ment avec les enfants
=
Mifandainga V *ny ankizy* N_1 *sy* Co *i Soa* N_0
Se mentent *V* les enfants N_1 et *Co* Soa N_0
Les enfants et Soa se mentent entre eux

Il est clair que, vu les différences de structuration ou d'agencement des groupes nominaux qui opposent formellement et sémantiquement les phrases factitives et réciproques aux phrases transitives et intransitives, seule une étude systématique et codée des transformations mises en jeu permettra d'envisager un quelconque traitement automatique ou assisté par ordinateur de la traduction franco-malgache.

Les diathèses dérivées

Il existe autant de diathèses dérivées que de types de compléments focalisables en sujet grammatical. Nous en avons dénombré sept, à ce jour, le septième étant constitué par les circonstanciels, dont la focalisation reste problématique pour autant qu'il s'agisse de compléments « libres » par opposition aux objets, qui sont des compléments « essentiels ». Je distinguerai ainsi deux sortes de diathèses dérivées : les diathèses « objectives » et les diathèses « circonstancielles ». Nos exemples seront, autant que possible, des résultats de l'application des opérations de permutation, d'effacement et d'ajout sur les phrases actives déjà proposées, opérations que nous préciserons en cas de besoin.

Les diathèses objectives

Elles résultent des opérations formelles suivantes appliquées sur une phrase active : 1° permutation du *GN* sujet (=: N_0) – avec le *premier objet* direct ou indirect (=: N_1) à focaliser, dans une phrase à deux arguments ou plus, – avec le *second objet* indirect ou (rarement) direct (=: N_2) à focaliser et ensuite avec le premier, dans une phrase à trois arguments ; 2° effacement du préfixe d'actif *Pfx =: mx-* et son remplacement par un affixe de passif *Pfx =: a-, Sfx =: -ana, -ina* ou par un circonfixe de relatif *Cfx =: x- ...-ana* ; 3° insertion de la préposition enclitique *Prép =: E + -na* entre le *GN* sujet permuté et le verbe. Nous passerons en revue les six principales diathèses objectives que nous avons identifiées.

La diathèse passive. Quand l'objet direct est focalisé, la phrase est à la voix passive, sauf avec un petit nombre de verbes, qui exigent la voix relative. Nous avons les exemples suivants de structures :

mx-V N_1 N_0 = *(V(-ana, -ina), a-V, x-V-a)-na* N_0 N_1

Mamangy ny reniny i Soa
Soa visite sa mère
= *Vangi-a-n'i Soa ny reniny*
Sa mère est visitée par Soa

Mikapoka ny alika i Be
Be frappe le chien
= *Kapoh-i-n'i Be ny alika*
Le chien est frappé par Be

Mandrodana ny tamboho ny mpiasa
Les ouvriers démolissent le mur
= *A-roda-n'ny mpiasa ny tamboho*
Le mur est démoli par les ouvriers

Mianatra ny lesona i Soa
Soa apprend la leçon
= *I-anar-a-n'i Soa ny lesona*
La leçon est apprise par Soa

On remarque que la phrase active à la diathèse factitive peut aussi prendre une forme passive à la diathèse passive ; c'est l'agent causé (N_0) qui se comporte alors comme un objet direct et l'agent causateur (N_e) comme un sujet agent, le verbe continuant à garder le préfixe complexe *Pfx =: amp-x-*. Les exemples suivants présentent alors les structures :

m-amp-x-V N_0 *(E, Prép* N_2*)* N_e = *amp-x-V-i-na* N_e *(E, Prép)* N_2 N_0

Mampi-antso an'i Soa i Be
Be fait appeler Soa
= *Ampi-antso-i-n'i Be i Soa*
Soa est faite appeler par Be

Mampandainga an'i Soa amin'ny ankizy i Be
Be fait mentir Soa aux enfants
= *Ampan-dainga-i-n'i Be amin'ny ankizy i Soa*
Soa est faite mentir aux enfants par Be

La diathèse instrumentale. Quand l'objet instrumental est focalisé, le verbe est à la voix passive à préfixe *Pfx =: a-* ou à la voix relative (voir Rabenilaina 1984). Mais le locuteur préfère cette dernière, qui est obligatoire en cas de phrase intransitive à transformer en phrase instrumentale. Les structures mises en relation sont les suivantes avec une phrase transitive :

$mx\text{-}V\ N_1\ (E + Prép)\ N_2\ N_0 = (a\text{-}V + x\text{-}V\text{-}a)\text{-}na\ N_0\ N_1\ N_2$

Exemple :

Mandidy ny hena (E + amin'ny) antsy pika i Soa
Soa coupe la viande avec (un + le) canif
= *(A-didi + an-didi-a)-n'i Soa ny hena ny antsy pika*
Le canif est-l'instrument-avec-lequel Soa coupe la viande

Elles se présentent comme suit avec une phrase intransitive (voir Rabenilaina 1985b) :

$mx\text{-}V\ (E + Prép)\ N_1\ N_0 = x\text{-}V\text{-}a\text{-}na\ N_0\ N_1$

Exemple :

Misarotro (E + amin'ny) tsihy i Be
Be se-protège-de-la-pluie avec (une + la natte)
= *I-sarotro-a-n'i Be (E + ny tsihy)*
(Une + la natte) est-l'instrument-avec-lequel Be se-protège-de-la-pluie

Comment faire comprendre à l'ordinateur que la phrase instrumentale intransitive est exclusivement à la voix relative par opposition à la phrase instrumentale transitive, qui peut être aussi à la voix passive à préfixe *Pfx =: a-* et que, de part et d'autre, la préposition *Prép =: am* est obligatoirement effacée ?

La diathèse destinataire. Si les données sont relativement simples en ce qui concerne la focalisation de l'objet destinataire dans une phrase intransitive, elles présentent une certaine complexité dans une phrase transitive. Les structures en relation sont les suivantes dans le premier cas, avec effacement obligatoire de *Prép* dans la phrase destinataire :

$mx\text{-}V\ Prép\ N_1\ N_0 = (V\text{-}i + x\text{-}V\text{-}a)\text{-}na\ N_0\ N_1$

Exemple :

Mandainga amin'i Be i Soa
Be ment à Soa
= *(Lainga-i + an-daing-a)-n'i Soa (E + *amin')i Be*
Be est-la-personne-à-qui Soa ment

Dans le second cas, la compatibilité affixale varie selon que le verbe présente ou non un sens « propre » et un sens « figuré ». Lorsqu'il ne s'emploie qu'au sens propre, les structures en relation sont les suivantes, avec effacement facultatif de la préposition devant le destinataire focalisé, dont le trait sémantique peut être humain ou non humain :

$mx\text{-}V\ N_1\ Prép\ N_2\ N_0 = (V\text{-}a + x\text{-}V\text{-}a)\text{-}na\ N_0 N_1\ (E + Prép)\ N_2$

Exemples :

Mamafy rano (E + eny) (amy + an-) tokotany ny i Soa
Soa asperge de l'eau sur la cour
=
(Fafaz-a + am-afaz-a)-n'i Soa rano (ny + Prép) tokotany
La cour est-l'endroit-sur-lequel Soa asperge de l'eau

Mamafy rano (E + eny) amin'i Be i Soa
Soa asperge de l'eau sur Be
=
(Fafaz-a + am-afaz-a)-n'i Soa rano (E + Prép) i Be
Be est-la-personne-sur-qui Soa asperge de l'eau

Lorsque le verbe s'emploie aussi bien au sens propre qu'au sens figuré, c'est-à-dire, en fait, lorsqu'il renferme deux verbes, les traits sémantiques de la tête du groupe prépositionnel destinataire entrent en jeu : si le destinataire est non humain (généralement un nom désignant un lieu de destination), il peut garder la préposition et la phrase où il est focalisé est facultativement à la voix passive ou relative ; exemple :

Mametaka sary (E + eo) amin'ny rindrina i Be
Be colle des photos sur le mur
=
(Petah-a + am-etah-a)-n'i Be sary (E + Prép) ny rindrina
Le mur est-l'endroit-sur-lequel Be colle des photos

Si le destinataire est, par contre, un humain, il perd la préposition et la phrase, où il est focalisé, se met exclusivement à la voix passive ; exemple :

Mametaka tehamaina amin'i Soa i Be
Be flanque une gifle à Soa
=
*(Petah-a + *am-etah-a)-n'i Be tehamaina (E + *Prép) i Soa*
Soa est-la-personne-à-qui Be flanque une gifle

On voit bien que la reconnaissance des phrases à la diathèse destinataire n'est pas simple en traduction franco-malgache : il faut identifier préalablement, au niveau de l'active, la valeur fonctionnelle liée au trait sémantique du complément focalisé, de même que l'emploi propre et/ou figuré du verbe. Les trois autres diathèses qui suivent présentent, relativement, moins de problèmes.

La diathèse dative. Lorsque l'objet datif est focalisé, la phrase ainsi obtenue se met à la voix relative avec tous les verbes, moyennant l'effacement obligatoire de la préposition *Prép =: am.* Les structures en relation sont, cependant, différentes selon

qu'on a affaire à une phrase intransitive ou à une phrase transitive. Elles se présentent comme suit dans le premier cas :

mx-V Prép N_1 N_0 = *x-V-a-na* N_0 N_1

Exemple :

Miresaka amin'i Soa i Be
Be parle à Soa
= *I-resah-a-n'i Be i Soa*
Soa est-la-personne-à-qui Be parle

La formule est la suivante dans le second cas :

mx-V N_1 *Prép* N_2 N_0 = *x-V-a-na* N_0 N_1 N_2

Exemple :

Manafina ny vaovao amin'i Soa i Be
Be cache la nouvelle à Soa
= *An-afen-a-n'i Be ny vaovao i Soa*
Soa est-la-personne-à-qui Be cache la nouvelle

La diathèse locative. La focalisation de l'objet locatif exige aussi que la phrase dérivée soit à la voix relative. Mais la présence de la préposition *Prép =: (E + Loc) (am + an)* est facultative. La relation d'équivalence est :

mx-V Prép N_1 N_0 = *x-V-a-na* N_0 *(E + Prép)* N_1

quand la phrase est intransitive ; exemple :

Monina (E + eny) (amy + an-) lohasaha ry Be
Les Be habitent dans (une + la) vallée
=
Onen-a-n-dry Be (ny + (E + eny) (amy + an-)) lohasaha
(E + (à + dans)) la vallée est-l'endroit-où-habitent les Be

Elle est :

mx-V N_1 *Prép* N_2 N_O = *x-V-a-na* N_0 N_1 *(E + Prép)* N_2

quand la phrase est transitive ; exemple :

Manafina vary (E + ao) (amin'ny + an-) hady i Be
Be cache du riz dans la fosse
= *An-afen-a-n'i Be vary (ny + (E + ao) (amin'ny + an-)) hady*
(E + dans) la fosse est-l'endroit-où Be cache du riz

Nous n'avons pas tenu compte de la possibilité d'occurrence de *Loc* à droite du groupe prépositionnel locatif aussi bien à l'actif qu'au relatif. Il faudra le faire le mo-

ment venu. Le caractère plutôt bizarre de la phrase locative où le même *GP* ne comporte pas de *Loc* à gauche de *Prép =: an* n'a pas été non plus pris en considération.

La diathèse de moyen. Lorsque l'objet de moyen est focalisé, deux types de structures équivalentes sont aussi en présence, selon qu'il s'agit de phrase intransitive ou de phrase transitive. La préposition *Prép =: (E + am (E + al))* est effaçable quand le complément (N_1 ou N_2) ferme la phrase relative à la diathèse de moyen, mais présent quand ce complément est extrait, c'est-à-dire préposé au verbe par la copule *Cop =: no + ro*. Nous avons la formule suivante en cas de phrase intransitive :

mx-V Prép N_1 N_0
= *x-V-a-na* N_0 N_1
= *(E + Prép)* N_1 *Cop-x-V-a-na* N_0

Exemple :

Manarivo (E + amin'ny (E + alalan'ny)) omby Rabe
Rabe s'enrichit (avec + au moyen de) les zébus
=
An-arivo-a-n-dRabe ny omby
Les zébus sont-le-moyen-avec-lequel Rabe s'enrichit
=
(Ny + amin'ny (E + alalan'ny)) omby (no + ro) an-arivo-a-n-dRabe
C'est (avec + par le moyen de) les zébus que Rabe s'enrichit

La relation est la suivante en cas de phrase transitive :

mx-V N_1 *Prép* N_2 N_0
= *x-V-a-na* N_0 N_1 N_2
= *(E + Prép)* N_2 *Cop x-V-na* N_0 N_1

Exemple :

Mandio ny fitaratra (E + amin'ny (E + alalan'ny)) ranon-tsavony i Soa
Soa nettoie la vitre (avec + au moyen de) le savon liquide
=
An-diov-a-n'i Soa ny fitaratra ny ranon-tsavony
Le savon liquide est-le-moyen-avec-lequel Soa nettoie la vitre
=
(Ny + amin'ny (E + alalan'ny)) ranon-tsavony (no + ro)
an-diov-a-n'i Soa ny fitaratra
C'est (avec + par le moyen de) le savon liquide que Soa nettoie la vitre

Outre le recours possible à la locution prépositive *Prép =: am al,* une autre façon de discerner le moyen par rapport à l'instrument, dans une phrase transitive, est l'applicabilité de l'opérateur causatif *Op =: manao* (faire, rendre) au verbe actif. Celui-ci perd alors son affixe *Pfx =: mx-* pour prendre une forme nominale ou adjectivale, selon les items. Dans l'exemple ci-dessus, le verbe actif commute avec l'adjectif *Adj =: madio* (propre) ; soit :

Manao madio ny fitaratra (E + amin'ny (E + alalan'ny)) ranon-tsavony i Soa
Soa rend propre la vitre (avec + au moyen de) le savon liquide

La diathèse circonstancielle

Quelle que soit la portée du complément circonstanciel (noté $N_{i, j}$..., par opposition aux arguments du verbe, à indice numérique) sur la phrase, sa focalisation exige que celle-ci soit exclusivement à la voix relative. Deux cas sont à envisager : le circonstanciel doit être extrait ou peut être extrait.

Le circonstanciel doit être extrait. Quand le circonstanciel focalisé exprime la manière (*Prép =: am + an* non suivi de *Dét*) ou la cause (*Prép =: (ami + noho)na*)), il est nécessairement extrait par la copule *Cop =: no + ro* et doit garder sa préposition, d'après la formule :

mx-V (Prép N_1, N_1 Prép N_2) Prép N_i N_0
= **x-V-a-na N_0 (Prép N_1, N_1 Prép N_2) Prép N_i*
= *(*E + Prép) N_i (no + ro)-x-V-a-na N_0 (Prép N_1, N_1 Prép N_2)*

Exemples :

Miandry an-dRamose (amim- + am-) pilaminana ny ankizy
Les enfants attendent le maître dans le calme
=
**I-andras-a-n'ny ankizy an-dRamose (ny + (amim- + am-)) filaminana*
(E + dans) le calme est-la-manière-avec-laquelle les enfants attendent le maître
=
*(*Ny + (amim- + am-)) filaminana (no + ro) i-andras-a-n'ny ankizy an-dRamose*
C'est (E + dans) le calme que les enfants attendent le maître

Mialokaloka eo am-body hazo nohon'ny hafanana i Soa
Soa se-met-à-l'ombre au pied de l'arbre à cause de la chaleur
=
**I-alokalof-a-n'i Soa eo am-body hazo (E + nohon') ny hafanana*
(E + à cause de) la chaleur est pour quoi Soa se-met-à-l'ombre au pied de l'arbre
=
*(*E + nohon')ny hafanana (no + ro) i-alokalof-a-n'i Soa eo am-body hazo*
C'est (E + à cause de) la chaleur que Soa se-met-à-l'ombre au pied de l'arbre

Le circonstanciel peut être extrait. Quand le circonstanciel focalisé désigne un lieu ou un moment, son extraction est facultative ainsi que la présence de la préposition *Prép =: (E + Loc) (am +an), E + am,* d'après la formule :

mx-V (Prép N_1, N_1 Prép N_2) Prép N_i N_0
= *x-V-a-na N_0 (Prép N_1, N_1 Prép N_2) (E + Prép) N_i*
= *(E + Prép) N_i (no + ro) x-V-a-na N_0 (Prép N_1, N_1 Prép N_2)*

Exemples :

Miresaka amin'i Soa ao (amin'ny + an') efitra i Be
Be s'entretient avec Soa dans la pièce
=
I-resah-a-n'i Be amin'i Soa (ny + ao (amin'ny + an')) efitra
(E + dans) la pièce est-l'endroit-où Be s'entretient avec Soa
=
(Ny + ao (amin'ny + an')) efitra (no + ro) i-resah-a-n'i Be amin'i Soa
C'est (la pièce l'endroit où + dans la pièce que) Be s'entretient avec Soa

Miresaka ny vaovao amin'i Soa ao (amin'ny + an') efitra i Be
Be expose la nouvelle à Soa dans la pièce
=
I-resah-a-n'i Be ny vaovao amin'i Soa (ny + ao (amin'ny + an')) efitra
(E + dans) la pièce est-l'endroit-où Be expose la nouvelle à Soa
=
(Ny + ao (amin'ny + an')) efitra (no + ro) i-resah-a-n'i Be amin'i Soa ny vaovao
C'est (la pièce l'endroit où + dans la pièce que) Be expose la nouvelle à Soa

Nous ne nous étendrons pas davantage sur la diathèse circonstancielle. Elle pose encore des problèmes autant théoriques que formels. Théoriquement, son existence n'est pas claire dans la mesure où un complément circonstanciel n'est pas, par définition, attaché au verbe, contrairement au complément d'objet. Formellement, sa position, dans la phrase active, n'est pas fixe, d'où la question de savoir comment aider la machine à l'identifier dans une phrase quelconque.

Remarques finales

Elles sont de trois ordres et portent sur la phrase non verbale, la préposition passe-partout *Prép =: am* et l'ordre habituel des arguments dans la phrase. Elles posent des problèmes d'interprétation et de représentation qu'il faut résoudre pour la machine.

La phrase non verbale ne comporte, en malgache, ni auxiliaire ni autre support, quand le prédicat est un adjectif (*Adj*) ou un substantif concret (*Nconcret*). Le malgache est, dans ce sens, une langue à prédicat, par opposition au français, qui est une langue à verbe. Exemples :

Salama Adj *i Soa* N_0
En bonne santé *Adj* Soa N_0
Soa est en bonne santé

Biby Nconcret *ny alika* N_0
Animal *Nconcret* le chien N_0
Le chien est un animal

Le malgache dispose, avons-nous dit, de la préposition passe-partout *Prép =: am =: amy, amina,* qui est équivalente aux prépositions françaises telles que *Prép =: à, de, par, pour, dans, avec.* Nous avons vu que la façon de la désambiguïser est d'iden-

tifier la ou les variantes qui alternent avec elle dans chaque type de complément. Nous rappelons les suivantes :

E + am	=	avec	:	instrument
am	=	à	:	datif
am	=	à, pour	:	destinataire humain
(E + Loc) am	=	à, dans	:	destinataire non humain
am + -na	=	par, de	:	cause
(E + Loc) (am + an)	=	dans, vers	:	lieu
am (E + al)	=	avec, par	:	moyen
E + am	=	à, de	:	temps
am + an	=	avec, en	:	manière

L'ordre habituel des arguments dans une phrase malgache est :

Prédicat - Complément - Sujet

alors qu'en français il est :

Sujet - Prédicat - Complément

Cet ordre vaut aussi bien pour les actives que pour les passives et les relatives, avec cette particularité que c'est l'objet ou le circonstanciel qui occupe, dans ces dernières, la position du sujet, l'agent étant attaché au verbe par la préposition enclitique *Prép =: -na.* Il va sans dire que, dans une phrase active, le sujet et le complément circonstanciel sont préposables au prédicat par antéposition à l'aide de la copule *Cop =: E + dia,* et par extraction à l'aide de la copule *Cop =: no + ro.* Ce n'est pas le cas de l'objet direct ; quant à l'objet indirect, il est susceptible d'extraction par *no + ro* et non d'antéposition par *E + dia.*

Références

ANDRIANIERENANA, Clément-Luc (à paraître) : *Les phrases verbales à opérateur affixal,* thèse de doctorat, Université d'Antananarivo.

JAONARISAONA, Bertin (à paraître) : *Les verbes composés en malgache,* thèse de doctorat, Université d'Antananarivo.

RABENILAINA, Roger-Bruno (1984) : « Verbes à instrumental et destinataire en malgache », *Linguisticae Investigationes,* VIII:1, Amsterdam, John Benjamins.

RABENILAINA, Roger-Bruno (1987) : *Lexique-grammaire du malgache. Constructions transitives et intransitives,* thèse de doctorat d'État, Université Paris 7, FO.FI.PA.

RABENILAINA, Roger-Bruno (1985b) : « Intransitifs à instrumental », *Hiratra,* 4, Antananarivo, SNIC.

RABENILAINA, Roger-Bruno (1990) : « Construction du dictionnaire électronique du malgache parallèlement à celui du français : le recensement des formes verbales », *Actes du Colloque. Les industries de la langue. Perspectives des années 1990,* tome I, Montréal, Office de la langue française et Société des traducteurs du Québec, pp. 253-262.

RABENILAINA, Roger-Bruno (1991a) : « Voix et diathèse en malgache », *Linguisticae Investigationes,* XV:2, Amsterdam, John Benjamins.

RABENILAINA, Roger-Bruno (1991b) : *Le verbe malgache,* Paris, AUPELF/UREF et Laboratoire de linguistique informatique, Université Paris 13.

RABENILAINA, Roger-Bruno (1993) : « L'intégration des différents parlers, signes manifestes de l'unicité de la langue malgache », *Language – a Doorway Between Human Cultures. Tributes to Dr. Otto Chr. Dahl on his Ninetieth Birthday,* Oslo, Novus Forlag.

RAHARINIRINA-RABAOVOLOLONA, Lucie (1991) : *Lexique-grammaire des composés du malgache. Les adverbes de temps,* thèse de doctorat, Université Paris 7.

RALALAOHERIVONY, Baholisoa Simone (à paraître) : *Les constructions adjectivales en malgache,* thèse de doctorat, Université Paris 7.

RANAIVOSON, Jeannot-Fils (à paraître) : *Le verbe support « manao » (faire),* thèse de doctorat, Université d'Antananarivo.

RASOAMALALAVAO, Claire (à paraître) : *Étude des déterminants et de leur expansion au niveau du groupe nominal objet direct du verbe,* thèse de doctorat, Université Paris 7.

RASOAZANANIVO, Florine (à paraître) : *L'enseignement de la langue maternelle dans le premier cycle du secondaire. Le cas du malgache,* thèse de doctorat, Université Paris 13.

RAZAFIMAMONJY, Jean-Paulin (à paraître) : *Les constructions symétriques et réciproques en malgache,* thèse de doctorat, Université d'Antananarivo.

33

La terminotique aux Services de traduction de Services gouvernementaux Canada

Jean QUIRION*

Services gouvernementaux Canada, Ottawa, Canada

• *Abstract* •

The field of computer-assisted terminology is booming. The Translation Services of Government Services Canada provides its terminologists with state-of-the-art software: TERMIUM, PUBLICIEL, and LATTER. TERMIUM is a linguistic data bank ; PUBLICIEL is a software program developed to publish terminology bulletins ; LATTER is a terminologist's workstation.

This paper first describes the features of these three work tools by an examination of the pros and cons of developing a software program in-house. The author then looks at the impact of computerization on terminology. Lastly, the author presents a list of researches currently underway in the field of terminology software development.

Introduction

L'informatique et la terminologie se côtoient depuis plusieurs années déjà. Les banques de données informatisées font depuis longtemps partie du paysage et les logiciels de gestion de données terminologiques sur micro-ordinateur circulent en grand nombre[1]. La recherche du logiciel idéal pour les besoins propres aux Services de traduction de Services gouvernementaux Canada s'est cependant révélée vaine. Le Ministère s'est alors résolu à la création des outils terminotiques désirés.

* Nous tenons à remercier Monique C. Cormier, Marie-Claude L'Homme, Ingrid Meyer et Gilbert Dupuis d'avoir enrichi de leurs commentaires la présente communication.

1. Voir Blanchon (1991).

Ce sont les caractéristiques de ces logiciels que la présente communication veut d'abord exposer. Nous aborderons ensuite les étapes ayant mené à leur création et discuterons des mérites de l'approche choisie. Nous poursuivrons en présentant l'incidence sur la chaîne de travail terminologique de l'avènement de la terminotique. En fin, nous traiterons de la nature des recherches en cours pour la poursuite du développement des applications.

Le mandat des Services de traduction de Services gouvernementaux Canada inclut la normalisation et la diffusion de la terminologie en usage dans la fonction publique fédérale et ailleurs au Canada[2]. Les moyens retenus à cette fin sont principalement la banque de données linguistiques TERMIUM et la publication d'ouvrages terminographiques.

Annuellement, la cinquantaine de terminologues des Services de traduction effectue environ 115 000 créations, mises à jour et annulations de fiches, et publie une douzaine de lexiques et vocabulaires.

L'importance de l'entreprise justifie donc à elle seule le recours à l'informatisation. Cette dernière a pris forme au début des années 70, soit bien avant la naissance du terme « terminotique » ! Ce qui était alors le Secrétariat d'État du Canada faisait l'acquisition de la banque de terminologie de l'Université de Montréal. En effet, le « UM » de « TERMIUM » provient de l'abréviation de « Université de Montréal ». Plus récemment, le Ministère créait PUBLICIEL, logiciel de préparation de publications terminographiques ainsi que LATTER, poste de travail du terminologue.

L'objectif était triple : réduire la saisie et la manipulation multiples des données afin d'améliorer la quantité et la qualité de la production.

Voyons brièvement en quoi consistent ces trois logiciels terminotiques.

TERMIUM

Banque de données relationnelles, TERMIUM, de par son million de fiches tirées de 55 000 sources, couvre à peu près tous les domaines du savoir humain. TERMIUM tourne sous le logiciel BASIS, sur gros ordinateur VAX.

En 1985, TERMIUM devenait accessible au grand public, par télécommunications. En outre, depuis 1990, TERMIUM est vendu sur disque optique compact (CD-ROM). Les langagiers, au pays et internationalement, constituent le gros de ses abonnés.

TERMIUM accueille, depuis peu, la terminologie rédigée en italien, néerlandais, suédois, norvégien et portugais, en plus du français, de l'anglais, de l'allemand et de l'espagnol habituels[3].

Les tiroirs de la banque constituent une autre nouveauté. Un tiroir est une partie vierge du système TERMIUM, louée ou vendue à des organisations pour la consignation et la gestion de leurs données terminologiques. Les organisations peuvent ainsi

2. Récemment réaffirmé dans Collet (1993).
3. Voir aussi Joe *et al.* (1992).

bénéficier de l'infrastructure mise en place par Services gouvernementaux Canada pour gérer leur terminologie, laissant le soin au Ministère de procéder à la maintenance du système, aux mises à niveau du logiciel, à la gestion de l'espace-disque, aux achats de matériel, aux copies de sécurité, etc.

PUBLICIEL

Conçu en 1990, PUBLICIEL simplifie la publication des ouvrages terminographiques. Le tri alphabétique selon les règles terminographiques, l'inversion des colonnes et les renvois aux synonymes figurent parmi les possibilités de mise en pages.

Le logiciel présente un écran de saisie dépouillé, où ne figurent que les champs des vedettes et des justifications[4] pour le français, l'anglais et, dans le cas des nomenclatures, le latin. La prochaine version de PUBLICIEL pourra traiter une langue supplémentaire, l'espagnol.

Cette prochaine version, en chantier, exploitera davantage une des caractéristiques actuelles de PUBLICIEL : la génération de lexiques et vocabulaires dans un format compatible avec le logiciel de traitement de texte WordPerfect. En effet, la mise en pages des entrées d'un vocabulaire de langue, spécialisée ou générale, obéit à un patron fort répétitif ; elle se destine donc bien à l'informatisation. Une fois les informations saisies, l'utilisateur coche à l'écran les options de mise en pages désirées : publication avec ou sans justifications, anglais-français ou français-anglais, avec ou sans renvoi automatique aux synonymes, etc. PUBLICIEL se charge ensuite de traiter les informations de chaque fiche selon les paramètres choisis. Le résultat du traitement est livré quelques minutes plus tard sous la forme d'un fichier WordPerfect.

Les informations terminologiques ainsi mises en page, WordPerfect est utilisé pour préparer le prêt-à-photographier. Cette dernière étape, auparavant confiée à des sous-traitants, rendait malaisés les changements de dernière minute ; son exécution à l'interne économise à la fois temps et argent. Ces avantages sont cependant obtenus au prix d'une période d'adaptation relativement longue ; ils ont exigé la formation des employés à de nouveaux logiciels et à de nouvelles méthodes de travail.

Conçu à partir du système de gestion de base de données Clipper, PUBLICIEL tourne sous DOS, dans l'environnement IBM, en version monoposte ou réseau local.

Si PUBLICIEL permet la saisie de certaines données terminologiques, c'est dans un but d'édition. Les terminologues ont vite confirmé ses limites quant à la gestion de la terminologie et réclamé un véritable poste de travail du terminologue.

LATTER

LATTER est l'acronyme de *L'ATelier du TERminologue*. Présentement simple logiciel de gestion de données terminologiques, il deviendra un véritable poste de travail du terminologue quand il intégrera un module de dépouillement assisté par ordinateur, une connexion directe avec TERMIUM, les fonctions de PUBLICIEL, etc.

4. Le terme recouvre les définitions, contextes, exemples et observations.

Contrairement à ce dernier, le poste de travail du terminologue possède tous les champs nécessaires pour le traitement complet d'une fiche terminologique basée sur le modèle de la fiche TERMIUM (domaines, paramètres, sources, codes d'auteur, dates de rédaction, etc.).

Livré au début de 1992, LATTER possède des champs de longueurs variables, caractéristique encore trop rare parmi les logiciels de ce type. En outre, la souplesse structurelle de sa fiche rend possible l'insertion à volonté de champs répétitifs. Ainsi, quand le terminologue désire ajouter un synonyme, une définition ou une langue à sa fiche, il appuie sur une touche et le champ se greffe à l'endroit désiré.

Grâce à son format d'exportation vers PUBLICIEL, LATTER bénéficie d'une grande richesse de présentation. Par ailleurs, l'interface TERMIUM-LATTER autorise un échange bidirectionnel de fiches, en vue de la mise à jour de TERMIUM.

De surcroît, l'interrogation en différé de TERMIUM représente une des options les plus intéressantes du poste de travail du terminologue. Cette fonction interroge automatiquement TERMIUM pour chacun des termes que le terminologue s'apprête à charger en banque ; ce dernier s'assure ainsi qu'il ne dupliquera pas une fiche existante.

La validation en direct signale au terminologue les données invalides. Une autre validation, en différé cette fois, repère les incohérences ou note l'absence d'informations essentielles, par exemple. Lors de la saisie, les informations répétitives peuvent être entrées sur une fiche-modèle. On peut aussi déplacer ou copier des blocs d'informations d'une fiche à l'autre.

Le poste de travail du terminologue et TERMIUM sont les piliers sur lesquels s'appuient les Services de traduction du gouvernement canadien pour gérer efficacement leurs données terminologiques.

LATTER, dont la commercialisation est à l'étude, tourne sur un micro-ordinateur IBM de type 386. Tout comme PUBLICIEL, il est programmé avec le logiciel Clipper.

Innovations riches en enseignements

Nous venons de décrire les logiciels terminotiques du Ministère ; abordons maintenant les étapes ayant mené à leur naissance. Elles sont riches en enseignements.

Tout d'abord, l'équipe de conception regroupe des participants de divers horizons : informaticiens, utilisateurs (préposés à l'édition ou terminologues) et terminologues-analystes (hybride entre le terminologue et l'informaticien).

Au départ, les utilisateurs exposent leurs besoins aux terminologues-analystes. Ceux-ci prennent le relais et travaillent de concert avec les informaticiens à la réalisation du produit. Régulièrement, les travaux sont présentés aux utilisateurs, qui corrigent alors toute erreur de trajectoire.

À la livraison du produit, on pourrait s'attendre à ce qu'il soit conforme au désir des utilisateurs. Or, il n'en est rien et les raisons en sont variées.

D'abord, les utilisateurs sont habitués aux logiciels commerciaux, souples et robustes. Or, la résistance naturelle au changement, alliée aux bogues et limites propres à tout nouveau logiciel, font hésiter l'utilisateur. Ce dernier s'étonne de ne pas retrouver toutes les fonctions utilisées dans d'autres logiciels ; il exige par exemple tous les raffinements du traitement de texte dans une base de données où, par la force des choses, les options d'édition sont minimales.

À cause de ces attentes élevées chez les utilisateurs, le logiciel a été plus ou moins bien accueilli. Un parrainage intense des utilisateurs tend à rétablir les choses, mais surtaxe l'équipe de formation et de dépannage. Par conséquent, la diffusion du poste de travail du terminologue s'effectue parcimonieusement, ce qui est regrettable.

Toutefois, un ingénieux maillage s'instaure entre utilisateurs novices et expérimentés pour la formation et le dépannage. Certains utilisateurs se chargent même des tests précédant la diffusion d'une nouvelle version.

Pour assurer un développement harmonieux de l'application, un groupe d'utilisateurs a été formé. Ce groupe rencontre mensuellement l'équipe de développement du logiciel ; les utilisateurs donnent la priorité aux activités de développement, en plus de proposer des améliorations, d'échanger des trucs, etc. La participation des utilisateurs à l'évolution d'un produit est naturelle et souhaitable.

Remarquable, l'influence réciproque entre l'utilisateur et le produit ne saurait être sous-estimée : l'arrivée d'un logiciel spécialisé renouvelle jusqu'aux méthodes de travail. Il arrive en effet que le poste de travail se révèle inapplicable pour une tâche imprévue ou, au contraire, qu'il favorise une nouvelle méthodologie. À cet égard, les réunions du groupe d'utilisateurs donnent souvent lieu à des débats passionnés.

Prenons par exemple l'échange de fiches entre PUBLICIEL, LATTER et TERMIUM. Les fiches créées avec LATTER en vue d'une publication sont maintenant chargées à TERMIUM au fur et à mesure, ce qui les rend interrogeables immédiatement. Auparavant, en prévision d'éventuelles mises à jour, certains terminologues préféraient conserver ces fiches à leur bureau jusqu'à ce que la publication soit terminée.

La chaîne de travail terminotique est-elle solide ?

Comment tous ces outils s'intègrent-ils dans la chaîne de travail terminologique des Services de traduction ? Tiennent-ils leurs promesses ? À cet égard, une synthèse des forces et des faiblesses de la terminotique aux Services de traduction est révélatrice. Après quelques généralités, nous aborderons la question par a) le dépouillement, b) la rédaction et la synthèse, c) le chargement et d) la diffusion et la mise à jour.

Certes, la convivialité, la souplesse et la puissance des logiciels créés représentent d'indéniables avantages. Dans leur sillage, la réduction des tâches répétitives apporte une satisfaction accrue aux langagiers, qui se consacrent à des activités faisant davantage appel à leur jugement. Le temps gagné, accordé aux travaux terminologiques proprement dits, favorise un produit de meilleure qualité.

Cependant, une des difficultés rencontrées jusqu'ici réside dans les délais d'amélioration des produits. Les efforts de recherche et de développement sont immenses et

les sommes engagées le sont tout autant. Comme ailleurs, les budgets de recherche sont régulièrement soumis aux coupures et aux gels. Il en résulte des délais d'amélioration difficilement tolérables pour des produits par définition imparfaits. Ce sont les ressources humaines qui sont alors appelées à compenser les imperfections du logiciel : rédaction d'aide-mémoires identifiant les circonstances problématiques de telle ou telle fonction, suivis de formations intenses, dépannages multiples, etc.

Au surplus, l'utilisation parallèle de plusieurs logiciels nuit à l'intégration des méthodes de travail. Ainsi, la mise à niveau de TERMIUM oblige la reprogrammation non seulement de son interface avec LATTER, mais aussi avec WordPerfect. Cette disparité favorise l'éparpillement des ressources de développement.

Dépouillement

Outre l'utilisation des macro-instructions exposée plus loin, le dépouillement se fait encore largement de la plus traditionnelle des façons : par la lecture des ouvrages par le terminologue. Par contre, le dépouillement assisté par ordinateur offre sans contredit le choix des meilleurs contextes disponibles et réduit la saisie, comme nous le verrons.

Rédaction et synthèse

La rédaction et la synthèse se font maintenant presque exclusivement à l'aide de trois logiciels : PUBLICIEL, LATTER et WordPerfect. La multiplication des logiciels tient de plusieurs facteurs, dont le rythme d'acquisition du matériel informatique et le rythme de développement des logiciels terminotiques au Ministère.

PUBLICIEL se retrouve chez les terminologues préparant une publication, en attendant un micro-ordinateur suffisamment puissant pour accueillir le poste de travail du terminologue. La rédaction des fiches se limite alors aux vedettes et justifications, qui constituent néanmoins l'essentiel de la fiche. Les informations complémentaires étant majoritairement répétitives (code d'auteur, domaine, date de rédaction, etc.), elles seront ajoutées au moment de la conversion à LATTER ou à WordPerfect, en route pour TERMIUM. Cela exige néanmoins des opérations supplémentaires de la part du terminologue ou de l'équipe de développement.

LATTER se retrouve progressivement chez les terminologues équipés du matériel adéquat. Moulé aux méthodes de travail du Ministère, il convient parfaitement, nous l'avons évoqué, à la synthèse et à la rédaction. La validation en direct ou en différé des fiches réduit particulièrement les risques d'erreurs.

Quant à WordPerfect, utilisé par le reste des terminologues, des macro-instructions amènent à l'écran un bordereau de saisie simple, évoquant celui de TERMIUM, où seront consignées les données.

Chargement

Les fiches créées à l'aide de PUBLICIEL, nous l'avons vu précédemment, sont d'abord converties, puis exportées vers LATTER ou WordPerfect, où elles sont complétées à l'aide de la fonction d'ajout global.

LATTER et WordPerfect possèdent tous deux une structure de fichiers reconnue par TERMIUM. Une série de macro-instructions transforme les fiches en format WordPerfect pour les rendre acceptables par TERMIUM, tandis que celles de LATTER sont directement assimilables par la banque.

Le poste de travail du terminologue a été conçu en gardant à l'esprit les nombreux échanges de terminologie qu'entretient le Ministère avec d'autres organisations, tant nationales qu'internationales[5]. Le format d'importation de LATTER apporte à cet égard une souplesse d'échange inégalée.

Le chargement à TERMIUM, organe ultime de diffusion terminologique au Ministère, se fait sur disquettes, en attendant l'avènement d'un réseau. L'absence d'un tel réseau diminue sensiblement l'efficience des interfaces entre les divers postes de travail du terminologue (à des fins de révision et d'échange), entre LATTER et PUBLICIEL, ainsi qu'entre LATTER et TERMIUM.

Diffusion et mise à jour

La diffusion de la terminologie, par TERMIUM ou par les publications, se trouve accélérée grâce à ces outils.

Depuis l'avènement de PUBLICIEL, la mise en pages et la préparation du prêt-à-photographier n'exigent que la moitié du temps requis auparavant. Quant on connaît la vitesse à laquelle évoluent les domaines de spécialité, les mois gagnés sont loin d'être négligeables.

Le contenu de TERMIUM se renouvelle plus rapidement, grâce aux interfaces avec LATTER et WordPerfect. Celles-ci éliminent la nécessité de saisie par des copistes, puis les relectures.

Recherches actuelles

Le développement des logiciels se poursuit. Les activités de recherche en cours portent surtout sur la phraséologie, le dépouillement assisté par ordinateur et l'intégration de PUBLICIEL à LATTER.

Les phraséologismes ont récemment acquis leurs lettres de noblesse auprès des terminologues[6]. Or, plusieurs questions se posent quand vient le moment de leur faire place sur la fiche terminologique. Doit-on créer un nouveau champ ou les ranger avec les exemples d'utilisation ? Un classement par catégories (nom + adjectif, nom + verbe, verbe + complément, etc.) serait sans doute idéal, mais ces catégories sont encore mal définies. Serait-il préférable d'indexer les phraséologismes avec les entrées ou de créer un index distinct ? Un tout récent séminaire du Réseau international de néologie et de terminologie sur le sujet a donné lieu à de vifs débats, sans toutefois amener de consensus.

5. Voir aussi Quirion (1992a).
6. Voir aussi Lainé *et al.* (1992).

Le besoin pressant d'informatisation du dépouillement a donné naissance à un programme de « dépouillement assisté par ordinateur », rédigé à partir du langage de macro-instructions de WordPerfect. Son utilisation est simple : le terminologue lit le texte à l'écran plutôt que sur papier ; au fil de sa lecture, il encadre de symboles précis les termes jugés intéressants. Une fois le dépouillement terminé, le terminologue lance une commande qui extrait du texte les termes marqués et les consigne dans un document distinct, en format d'exportation vers le poste de travail du terminologue. Au choix, le terminologue inclura ou non la phrase dans laquelle le terme apparaît. Le domaine, la source, certains paramètres, le code d'auteur, la date de rédaction sont ensuite ajoutés globalement avant l'importation dans LATTER.

En attendant des logiciels de dépouillement assisté par ordinateur suffisamment performants, cette méthode exploite une forme de dépouillement assisté qui satisfait les tenants d'une philosophie discutable, selon laquelle *tous* les textes doivent être lus par le terminologue et non par la machine. Les travaux d'informatisation de ce maillon vital de la chaîne doivent se poursuivre, afin d'allier qualité et exhaustivité.

L'intégration de la fonctionnalité de PUBLICIEL à LATTER est toute naturelle. La souplesse de rédaction offerte par LATTER, jumelée à la richesse de présentation de PUBLICIEL, justifient la fusion des deux outils (qui partagent d'ailleurs la même plate-forme logicielle) en un seul, plus polyvalent.

Enfin, les nombreuses améliorations suggérées par les utilisateurs complètent le programme des activités de développement.

Conclusion

Les Services de traduction canadiens ont misé sur le mariage entre l'informatique et la terminologie pour mener à bien leur mandat de diffusion et de normalisation. L'entreprise, pluriannuelle, est exigeante en ressources humaines et financières.

Nous espérons que les leçons de notre expérience serviront à ceux qui sont tentés par l'aventure de la création et du développement de produits terminotiques. Les outils présentés aujourd'hui se révèlent encore limités, mais leur place au cœur du travail terminologique est acquise. Le développement doit se poursuivre.

Références

AUGER, Pierre, DROUIN, Patrick et Marie-Claude L'HOMME (1991) : « Automatisation des procédures de travail en terminographie », *META*, vol. 36, nº 1, pp. 121-127.

BLANCHON, Élisabeth (1991) : « Choisir un logiciel de terminologie », *La Banque des mots, numéro spécial « Les logiciels de terminologie »*, nº 4, pp. 5-57.

COLLET, Roger (1993) : « La normalisation terminologique, une nécessité à l'heure de la mondialisation des marchés », *L'Actualité terminologique,* vol. 26, nº 1, p. 11.

JOE, Gregg, HUTCHESON, Helen et Christine LEONHARDT (1992) : « Multilateral Trade, Multilingual Terminology: New Directions! », *L'Actualité terminologique,* vol. 25, nº 4, pp. 9-11.

LAINÉ, Claude, PAVEL, Silvia et Monique BOILEAU (1992) : « La phraséologie – Nouvelle dimension de la recherche terminologique. Travaux du module canadien du RINT », *L'Actualité terminologique*, vol. 25, nº 3, pp. 5-9.

QUIRION, Jean (1992a) : « L'acquisition et l'échange de terminologie », *L'Actualité terminologique*, vol. 25, nº 1, pp. 17-18.

QUIRION, Jean (1992b) : « Trois larrons en foire: TERMIUM, PUBLICIEL et LATTER », *L'Actualité terminologique*, vol. 25, nº 3, pp. 12-14.

34

Bilan et prospectives

Jean-Claude Lejosne

Université de Metz, Metz, France

Le titre de cette contribution faite en clôture est tout à fait explicite. Selon le vœu des organisateurs, il s'agissait de dresser un bilan de la recherche dans le domaine retenu pour ces Troisièmes Journées scientifiques du réseau Lexicologie, Terminologie, Traduction de l'AUPELF-UREF consacrées à la traductique (traduction automatique et traduction assistée par ordinateur), à partir des communications soumises à l'attention des participants, et d'essayer de capter les perspectives qui s'en dégagent.

Introduction : vous dites « bilan » et « prospectives » ?

En tant que *bilan,* les pages qui suivent se présentent donc comme le récapitulatif des trois journées de travail. Elles sont fondées sur une écoute attentive des communications faites et une évaluation des tendances qui se dégagent dans le domaine étudié. Il s'agit en quelque sorte d'une photographie instantanée de reportage, qui présente tous les qualités et défauts de ce genre de support médiatique : le côté positif est évidemment la valeur de spontanéité et d'adhésion à la réalité du moment, le côté négatif est sa partialité, à la fois dans le sens objectif et subjectif du terme.

En tant que *prospective,* elles s'aventurent sur le terrain mouvant de la futurologie ; le rapporteur se place à cet égard sous le contrôle des autres participants aux Journées et précise que les perspectives qu'il dessine valent tout au plus le moyen terme, tant le rythme des changements semble rapide dans ce domaine des technologies de pointe : les photographies prises sur la situation en TA et en TAO sont d'autant plus rapidement frappées de caducité que l'évolution des choses est fortement influencée par les contingences économiques et que la prospective en la matière est encore plus délicate que dans le domaine de la recherche.

On notera que le rapporteur, qui n'a pas percé le secret de l'ubiquité, n'a pu intégrer dans ce bilan le fruit des travaux présentés dans la session du 02.10.93 organisée en parallèle ; les références à des exposés de cette session se fondent sur des notes remises par les intervenants ou sur les résumés publiés dans le livret-programme.

Bilan : TA et TAO : la troisième génération ?

Plusieurs communications se sont attachées à rappeler les grandes étapes de l'histoire de la TA et de la TAO et à rappeler les grandes lignes des deux générations de systèmes qui ont prévalu jusqu'à l'aube des années 90.

Il y a consensus pour reconnaître que la première génération, jusque vers 1976, a mis l'accent sur la définition des formalismes et la traduction directe. La deuxième génération a cherché plutôt à s'appuyer sur le transfert, avec plusieurs niveaux représentation correspondant grossièrement à l'analyse morphologique, l'analyse syntaxique et l'analyse sémantico-pragmatique, en amont de l'interface de transfert, et sur l'élaboration de diverses formes de langages-pivots plus ou moins abstraits et libérés des contraintes propres aux langues naturelles.

D'autres systèmes ont commencé à émerger au début des années 90, mais, à l'inverse de ce qui s'était passé jusqu'alors, il y a controverse sur l'évaluation de leur degré d'originalité et de nouveauté. En effet, il ressort des historiques faits sur la spécialité étudiée que, au tournant de la décennie en cours, après successivement la destitution du *grand manitou,* le règne éphémère du *petit manitou,* la tentative de relance de la part des *petits prêtres,* aucun grand pontife ou vizir n'a pu sortir des divers conclaves tenus par les spécialistes de linguistique et d'informatique travaillant en plus ou moins bonne intelligence.

Bien que la communication soit rendue facile par le miracle du courrier électronique qui fait voltiger messages et fichiers entre les rives du Pacifique et de l'Atlantique (tout en contournant le plus souvent l'Afrique), on peut se demander maintenant qui sera le vainqueur du début de polémique qui se dessine dans le monde de la traductique, qui s'imposera entre les tenants des modes fondées sur la statistique et les promoteurs du lexique et de la terminotique, quel rôle joueront les aventuriers de l'éditique et autres formes de formatique ? Bref, nous avons compris que, dans le paysage médiatique de la traductique de ces dernières années, il y a beaucoup de -iques (hics ?) auxquels on cherche à apporter, *hic et nunc*, des solutions qui semblent relever souvent de l'*ad hoc* (Haddock ?).

À parler plus sérieusement, on constate au moins qu'il n'y a pas consensus quant à l'avènement de la troisième génération. L'image donnée par l'évolution des choses en TA et en TAO en ce début des années 90 est une image floue (sans être confuse), avec le foisonnement habituel des systèmes, prototypes, maquettes, sans qu'on puisse cependant parler d'une percée impressionnante de la part d'un modèle original et appelé à s'imposer. Comme on le verra plus bas, le rapporteur serait tenté de parler d'un changement radical d'orientation, et donc de deuxième génération collatérale, plutôt que de génération nouvelle devant porter le numéro trois.

Le cadre économique : à l'heure de la crise

La lenteur de cet avènement, dont la réalité reste à démontrer, trouve partiellement son explication dans le durcissement des contraintes économiques.

Toujours plus vite, meilleur et moins cher

Le monde industriel est évidemment à la recherche du meilleur rapport qualité-prix. L'exigence de qualité se manifeste par l'affinement des modèles et procédures d'évaluation, ce qui passe par une définition plus serrée des critères applicables, tant du point de vue de l'efficacité que de celui de la traductologie. La recherche sur les possibilités d'automatiser l'analyse et d'élaborer des modules correcteurs y a également trouvé son compte.

Le critère de prix a surtout contribué au renforcement des efforts faits pour mettre au point de nouveaux outils appelés à être intégrés dans l'*environnement traductionnel,* (le PTT ou poste de travail du traducteur) et destinés à augmenter le rendement et l'efficacité du traducteur ou du réviseur, sans sacrifier la qualité et tout en respectant les lois de l'ergonomie. À en croire leurs promoteurs, certains outils semblent avoir déjà atteint un niveau de puissance tel que l'on peut exprimer une certaine inquiétude quant au droit de l'homme à l'erreur : à quand la machine qui prendra sa revanche et enverra, par exemple, une petite décharge électrique au malheureux besogneux qui, après une nuit sans sommeil, aura malencontreusement tapé *actually* pour traduire *actuellement* ? On taira la sanction appliquée en cas de récidive, pour épargner les âmes sensibles.

Autre forme de vengeance que pourrait exercer la machine si on essaie de la *court-circuiter,* à mesure que la confiance dans l'avenir de la TA diminue : les systèmes de génération multilingue de textes, à partir de bases de données complexes, selon la théorie de la relation sens <=> texte (Mel'čuk), avec renvoi à tous les scripts, scénarios, patrons et autres formes de gabarits (*templates*).

La crise, encore et toujours

En ces temps de crise dont on ne voit pas bien la fin, les bailleurs de fonds se font rares et exigeants. Dans le secteur public, la recherche académique continue à manifester son dynamisme légendaire et la coopération est d'autant meilleure qu'elle est pratiquement obligatoire, que ce soit pour des raisons d'investissement partagé ou parce que, comme c'est le cas pour les contrats octroyés par la Communauté européenne, la dimension internationale est un prérequis au dépôt d'une demande.

Les choses sont moins claires dans le secteur privé. Il apparaît que, en ces temps difficiles, les quelques fonds que les grandes sociétés spécialisées sont prêtes à dégager pour la recherche ne vont pas en priorité vers les recherches en TA ou TAO ; ils iront à la rigueur vers des projets ciblés en TALN ou en génie linguistique.

La crise a aussi, semble-t-il, élargi le fossé technologique entre le *Nord* et le *Sud.* Les collègues du monde africain francophone présents ont pu faire part de leurs dif-

ficultés. Il convient de leur rendre un hommage tout particulier : alors que les chercheurs des pays industrialisés continuent, tant bien que mal, à progresser sur une lancée remontant à au moins trois décennies, ces collègues font œuvre de pionniers dans des États dont les ressources financières et logistiques sont minces et menacées. Or c'est justement dans leurs pays que les besoins sont les plus grands et les plus urgents, ne serait-ce que pour sauver de nombreuses langues et cultures nationales de l'oubli ou pour propager certaines langues véhiculaires devenues supports d'enseignement et vecteurs d'une littérature et d'une dignité nouvelles. On saura à cette occasion saluer le rôle joué par des organismes tels que l'AUPELF ou la toute nouvelle Union européenne pour que s'opèrent les transferts de technologie et de connaissances nécessaires et que la promotion de la formation soit soutenue.

La prospective : où est/sera la nouveauté ?

Un environnement plus humain...

Un mot d'abord sur le contexte psychologique : c'est celui du lendemain de fête, celui de la fin des ambitions déraisonnables, celui de la modestie retrouvée. Et puis la priorité est redonnée au résultat, aux dépens de l'élégance de la méthode, de la procédure et de la solution à laquelle on a abouti. Il n'est plus honteux de déclarer que l'on n'exclut pas, dès le départ, la nécessité de procéder à une postédition soignée, tout en sachant que cela est un travail des plus fastidieux, souvent considéré comme une forme larvée de sanction, que l'on déguise sous le nom de promotion pour mieux faire accepter la corvée.

Cet aveu a donné une impulsion nouvelle aux travaux en recherche et développement pour la mise au point de divers outils relevant du génie logiciel et du traitement automatique des langues naturelles (TALN) et visant à améliorer le rendement de cette tâche. Après que les orthogiciels soient devenus monnaie courante dans des dizaines de langues à graphie alphabétique, les chercheurs et ingénieurs s'activent sur des formes de correcteurs plus élaborés – en particulier les correcteurs syntaxiques – en attendant des formes plus satisfaisantes de correcteurs sémantiques et stylistiques.

Dans le même ordre d'idée et toujours dans le sens de l'*humanisation* du travail du traducteur choisissant de (ou condamné à ?) travailler devant un écran, les spécialistes renforcent la part qu'il convient de faire à l'ergonomie en élevant le degré de convivialité des systèmes. Les recherches sur les modes d'interactivité et de dialogue homme/machine y trouvent largement leur compte et sont appelées à poursuivre leur développement.

En vision globale, cette évolution conduit donc à un relatif abandon de la TA proprement dite, au profit de diverses formes de TAO, ou plutôt à un effacement de la frontière entre les deux niveaux d'automatisation que représentent ces deux sigles.

Ce qui est aussi une autre façon de dire que, paradoxalement, les systèmes de la pseudo-troisième génération sont moins complexes – tant du point de vue des fondements linguistiques que de celui de la programmation informatique – que ceux de la deuxième.

Maudite ambiguïté...

Les qualités que devront présenter les systèmes en cours d'élaboration ou dans les cartons des concepteurs ont été rappelées au hasard des communications. Les qualificatifs utilisés trahissent la volonté pragmatique, à l'image des contraintes imposées par un marché difficile et une clientèle qui donne la priorité à l'efficacité et au rendement du système, qu'il s'agisse du secteur privé (les grandes sociétés ayant des volumes de traduction considérables) ou du secteur public (les institutions internationales, les offices chargés de l'application de la loi dans les États et nations multilingues). On attend donc des systèmes qu'ils soient *robustes* (sans que l'on sache très bien ce que ce vocable recouvre lorsqu'il s'applique à un système implanté sur une machine, depuis la résistance aux coups de marteau assénés par le traducteur cédant à la crise de nerfs, jusqu'à la capacité à s'accommoder des carences de l'esclave payé à la ligne et parvenu nuitamment à la 147e page d'un pensum sur la vie amoureuse de l'*hélix pommata* [l'escargot de Bourgogne, que l'on trouve surtout, comme son nom l'indique, dans les élevages de Hongrie] et *contrôlés par la tâche* (si on accepte cette formule pour traduire *task-driven*).

En d'autres termes, on reconnaît que le passage à l'implémentation ne doit pas être trop précoce, qu'il faut avoir effectué, avant d'y procéder, plus de travail en linguistique. C'est en quelque sorte la revanche des informaticiens sur les linguistes dans les équipes de recherche en TA ou TAO. Si tant est que la séparation entre les représentants de l'une et l'autre spécialité est possible, on sait que le ratio numérique idéal entre les premiers et les seconds se situerait du côté de 1/5 en version basse, 1/8 en version haute. C'est dire que le problème essentiel dans ce domaine reste celui de la désambiguïsation de l'énoncé en langue naturelle. Et la proclamation de l'informaticien affichée au-dessus de l'entrée de son bureau – « j'ai le droit de savoir exactement ce que les linguistes attendent exactement de moi » – garde toute sa force.

Il est donc universellement reconnu que la somme de travail à fournir pour percer la nature des processus cognitifs activés par le traducteur humain reste énorme. Même ceux qui prétendent avoir mis au point le système *robuste* dont il a été question plus haut finissent généralement par reconnaître qu'il leur reste de nombreux problèmes *durs à cuire* que même la méthode *shake-and-bake* ne saurait résoudre complètement. On citera en vrac la question de la portée de la négation et de l'interrogation, celle de l'anaphore et de la cataphore ou, plus généralement, de la coindexation et des dépendances à distance, celle de la détermination, celle de la quantification, et surtout celle de la présupposition et autres arcanes auxquelles les spécialistes de ce qu'il est convenu d'appeler l'intelligence artificielle se sont attaqués.

Puisque le passé n'a pas fait ses preuves et qu'un système d'inférence et de désambiguïsation prétendant émuler son équivalent humain n'est pas pour demain, il faut donc chercher ailleurs ou, plus simplement, mieux explorer certaines avenues que l'on avait quelque peu négligées à l'époque de l'euphorie.

Quelques options montantes...

Les communications faites lors de ces Journées ont dégagé plusieurs options généralement prometteuses.

L'option « lexicaliste »

Son succès est trahi par le nom même des projets : Genelex, Transterm... Où l'on voit avec plaisir le regain d'intérêt pour tous ces problèmes en relation avec l'identification des termes, la sémantique lexicale, la phraséologie et les idiomes, sans pour autant négliger ce que la syntaxe peut apporter à l'étude du sens. Une façon de faire disparaître la barrière entre les deux sciences et de remettre à l'honneur la notion de sémantaxe, par le biais de l'étude de toutes les formes lexicales complexes, qu'il s'agisse des groupes nominaux complexes, des syntagmes prépositionnels, etc.

Les méthodes quantitatives

Il faut probablement saluer aussi cette volonté d'exploiter plus largement que ces dernières années les potentialités offertes par la statistique linguistique et la théorie de l'information, en allant au delà des simples mécanismes de préférence et de pondération lorsque le système livre plus d'une lecture ou interprétation.

Ce mouvement a, à son tour, conduit à l'élaboration de divers outils allant de la moulinette à mots (autre traduction pour *word cruncher* ?) et du concordancier monolingue ou multilingue, aux programmes d'établissement des corrélations et des *comptes d'information mutuelle.*

L'exploitation des traductions existantes

L'idée a longtemps sommeillé dans l'esprit de nombreux chercheurs chevronnés issus du monde de la traduction humaine : comment mieux tirer parti des milliers de pages que l'on a produit au cours d'une carrière, comment exploiter les mémoires de traduction.

Ce n'est évidemment pas par hasard si ce sont des pays profondément attachés au multilinguisme comme le Canada ou la Suisse qui se sont placés à la tête du progrès dans ce domaine. Dans la mesure où les traductions humaines qui seront prises comme références ont été validées, le principe directeur a conduit à un développement spectaculaire des techniques d'alignement et des méthodes relevant de ce qu'il est maintenant convenu d'appeler la formatique et l'éditique ; le fait qu'une seule norme semble devoir s'imposer au niveau international pour le balisage des textes (SGML) devrait contribuer à renforcer cette tendance.

Plus de connaissances...

L'essor des sciences cognitives, presque en relais à l'*intelligence artificielle*, se manifeste également dans les projets de TA et TAO les plus récents par un intérêt renouvelé pour les problèmes de modes d'acquisition et de compréhension, et, plus encore, de représentation des connaissances pour la constitution des bases correspondantes.

Ce problème est lui-même mis en relation avec celui du mode de conceptualisation et donc, pour la TA(O), celui de la création des interlingues conceptuelles. La

recherche est, semble-t-il, en pleine effervescence sur ce point, l'accord se faisant sur le fait que les structures d'interface doivent se situer à un haut niveau d'abstraction et d'universalité des données.

... Et de textualité

Les deux tendances précédentes rejoignent celle qui consiste à intégrer, surtout au moment de l'analyse, des traits relevant de la textualité. Là encore, il s'agit de restituer au processus une part de sa dimension humaine et humaniste. Les congressistes ont entendu avec plaisir parler de force illocutoire, structure rhétorique, progression thématique, ordre et registre. La DRT a été remise à l'honneur, de même que les questions d'établissement de la structure communicative, en remettant en avant le calcul des structures thématiques, rhématique et phématiques...

En guise de conclusion

Aux deux questions centrales que l'on devait se poser, à savoir dans quel sens la communauté scientifique va investir ses efforts et la communauté économico-industrielle ses fonds en R & D, on répondra avec un optimisme raisonnable et raisonné, tout en restant dans le flou. Il ne peut de toute manière en être autrement, tant qu'il existera ce hiatus quasiment insurmontable entre les implications de ces deux questions. En effet, les délais demandés par les premiers (les chercheurs) sont souvent jugés inacceptables par les seconds (les bailleurs de fonds) qui voudraient que l'on trouve et découvre sur commande. On se souvient de ce collègue qui, rapportant sur le projet conduit par son équipe, a fait comprendre que, dans son laboratoire, à la fièvre stimulante du chercheur s'était complètement substituée l'angoisse paralysante précédant et suivant la tournée d'inspection effectuée par des évaluateurs eux-mêmes mis sous pression.

Il apparaît en tout cas que, dans le domaine clos de la TA et de la TAO, on ne soit pas appelé à assister au lancement de grands projets nouveaux aux ambitions révolutionnaires. En simplifiant grossièrement, nous assistons actuellement, tantôt au bouclage de ce qui a été mis en orbite dans un passé assez récent, tantôt à l'optimisation de l'existant, qu'il soit déjà commercialisé ou non. C'est dans ce dernier domaine que la communauté scientifique fait preuve du plus grand dynamisme. Ceci avant tout dans le sens de la TAO rebaptisée THAM (pour traduction humaine assistée par la machine – une nuance de taille !). En d'autres termes, on s'éloigne nettement de la TA proprement dite pour se tourner vers des programmes d'aide qui sont, eux, de plus en plus sophistiqués. Les conclusions des divers groupes d'évaluation tels que EAGLES ou l'European Association of Machine Translation sont tout à fait éloquentes à ce sujet.

D'autre part, il convient de dire que le programme, déjà fort chargé, n'a pas permis d'entrer dans le détail de toutes les retombées possibles que l'on peut attendre des travaux et recherches dont nous avons dessiné les tendances. On a cependant pu déceler que de nombreuses applications pourraient être développées, en particulier dans le domaine de l'EAO (enseignement, en particulier didactique des langues étrangères), la DAO (documentation, avec toutes les extensions telles que la constitution et

la consultation des banques de données multilingues), l'hypertexte et l'hypermedia multilingue, les plates-formes de communication multilingues, avec un langage plus ou moins contrôlé, etc. Or ce sont justement ces applications qui seraient utiles aux pays moins avancés en technologie.

En sa qualité de dernier orateur, le rapporteur soussigné a eu le privilège de remercier au nom des participants les organisateurs de ces Journées scientifiques. Cet honneur s'est doublé d'un immense plaisir : il était particulièrement agréable d'exprimer la gratitude de tous envers monsieur le professeur André Clas, ses collègues et son équipe pour la perfection de l'organisation, concernant tant le colloque lui-même que la vitrine technologique, pour la très haute qualité de la substance scientifique partagée et enfin pour quelque chose que la machine ne nous donnera jamais : un incomparable sens de l'hospitalité, une amabilité et une prévenance de tous les instants, le souci du bien-être de chacun. Un grand merci à tous les amis de la Belle Province.

Vitrine technologique

LINGUISTIC PRODUCTS
PC-TRANSLATOR

Le **PC-Translator** de *Linguistic Products* (P.O. Box 8263, The Woodlands, Texas 77387 USA) est un logiciel de traduction automatique pour compatibles IBM. L'extrême facilité d'enrichissement et d'adaptation de sa base de connaissances permet à l'utilisateur de lui apprendre à traduire selon ses préférences terminologiques et stylistiques. Des *jokers* simples ou multiples rationalisent aussi bien les dictionnaires que la grammaire. De plus, le texte d'arrivée conserve tous les codes typographiques du texte de départ. Vendu à moins de mille dollars US, **PC-Translator** est utilisé depuis 1985, dans l'une de ses diverses versions (de l'anglais vers le français, l'espagnol, l'italien, l'allemand, le danois, le suédois, le norvégien, le portugais et vice-versa), par des services internes et externes de traduction du monde entier appelés à produire des documents volumineux et répétitifs.

DOCUMENSA
EDIBASE

Si une bonne gestion de votre banque de données terminologiques est cruciale et que vous désirez en optimiser la consultation, le logiciel **EdiBase** est la solution. Logiciel de gestion et d'interrogation de bases de données terminologiques, **EdiBase**, malgré sa puissance, demeure un outil simple et convivial. Il facilite la production de lexiques, de glossaires, ou d'index de façon simple mais efficace. Pour éviter les erreurs coûteuses, **EdiBase** peut même valider automatiquement vos fiches terminologiques par un thésaurus de votre propre création. Les thésaurus créés avec **EdiBase** sont des outils d'aide à la recherche et à la saisie. **EdiBase** permet de travailler avec plusieurs thésaurus à la fois. On peut créer un thésaurus de codes de sources, un thésaurus d'auteurs, ou de réviseurs, ou comme il est le plus fréquent, un thésaurus des domaines. **EdiBase** se prête aussi très bien au recensement des ouvrages cités en référence sur les fiches terminologiques.

CENTRE D'INNOVATION EN TECHNOLOGIES DE L'INFORMATION(CITI)
TRANSSEARCH

Le système-prototype **TransSearch** est un concordancier bilingue qui met les traductions antérieures au service de la production en cours. Interrogé sur un problème de traduction particulier, **TransSearch** peut immédiatement en repérer toute instance présente dans sa « mémoire de traduction » et l'afficher avec, à son côté, la solution qui avait été élaborée à cette occasion. Ceci est rendu possible par une nouvelle technologie « d'appariement » de textes qui permet à **TransSearch** de reconstruire automatiquement les liens de traduction qui unissent les segments des textes de départ et d'arrivée.

CLC LIMITÉE
TERM CRUNCHER

Le logiciel de dépouillement terminologique **Term Cruncher** est un outil essentiel en terminotique. **Term Cruncher** est un logiciel qui lit intelligemment un fichier texte et recense presque tous les termes qui y sont contenus, ainsi que les collocations qui apparaissent de nombreuses fois. Le logiciel recherche automatiquement les équivalents dans votre base de données lexicales ou terminologiques.

LES LOGICIELS TRADULOG
PROTERM

Proterm est un logiciel de gestion de fiches terminologiques comprenant la gestion des fiches, la gestion des domaines et la gestion des sources (bibliothèque).

Proterm est composé de deux modules principaux : un module résidant et un module autonome. Le module résidant regroupe les fonctions concernant la mise à jour et la consultation des fiches. L'utilisateur peut rechercher un terme dans un champ particulier ou dans tous les champs de la fiche. **Proterm** inclut aussi deux fonctions de copier-coller : de **Proterm** vers le traitement de textes et vice-versa, toutes deux de 1716 caractères.

Le module autonome contient les fonctions du module résidant (sauf les fonctions de copier-coller) en plus des fonctions de gestion de domaines et de sources (bibliothèque), de conception de la structure de la fiche et de création de nouveaux fichiers.

SERVICES DE TRADUCTION DU GOUVERNEMENT DU CANADA
POSTE DE TRAVAIL DU TRADUCTEUR (PTT)

Le **PTT** est un outil qui permet au traducteur d'accéder simultanément à diverses fonctions susceptibles de faciliter ou d'accélérer son travail. Le **PTT** est construit autour d'un concept de base, soit l'exploitation de matériel et de logiciels courants permettant d'informatiser bon nombre de fonctions liées à l'acte de traduction. Au nombre des aides mises à la disposition du traducteur, mentionnons des bases de données terminologiques et textuelles, un logiciel de conjugaison et des dictionnaires électroniques.

SERVICES GOUVERNEMENTAUX DU CANADA
SERVICES DE TRADUCTION

La direction de la terminologie et des services linguistiques présente ses publi-

cations terminologiques et linguistiques et offre des démonstrations de la banque de données linguistiques du Canada, **TERMIUM sur CD-ROM.**

OFFICE DE LA LANGUE FRANÇAISE
BANQUE DE TERMINOLOGIE DU QUÉBEC

La Banque de terminologie de l'Office de la langue française (**BTQ**) est un dictionnaire bilingue informatisé qui compte plus de deux millions de termes scientifiques et techniques dans plus de 220 domaines. À partir d'un terminal, ou d'un micro-ordinateur, et d'un modem, vous accédez directement aux fiches terminologiques de la **BTQ**. L'abonnement annuel de 200 $ inclut le logiciel de communication pour les appareils IBM et compatibles, la formation au langage d'interrogation, les suivis et les rappels de formation, le bulletin d'information *Réseau-BTQ* et le dépannage le cas échéant.

La **BTQ** est accessible du lundi au vendredi de 7 h à 24 h, le samedi de 7 h à 17 h et le dimanche de 9 h à 17 h.

BYTETOWN CONSULTING
MERCURY/TERMEX (MTX^{MC})

Bytetown Consulting est le distributeur exclusif de **Mercury/Termex (MTX^{MC})** au Canada. **MTX** est un programme intégré de gestion de lexiques et de dictionnaires. Le système de gestion de terminologie **MTX** comprend toutes les fonctions nécessaires pour créer, gérer et interroger des bases de données lexicales ou terminologiques. Des lexiques et des dictionnaires précompilés **MTX** sont aussi disponibles. **MTX** a été adopté par le Bureau de la traduction du Secrétariat d'État. De plus, **MTX** est la base de données terminologiques la plus utilisée au Canada.

IBM DEUTSCHLAND
TRANSLATION MANAGER/2

IBM Translation Manager/2 : comment traduire efficacement.

TM/2 d'IBM est un outil de productivité conçu pour les professionnels de la traduction. Ses principales caractéristiques sont les suivantes :

* un éditeur de texte spécialement adapté à la traduction ;
* une mémoire de textes traduits ;
* un gestionnaire de fichiers terminologiques ;
* la capacité de traiter 19 langues de départ ;
* la protection des codes de formatage de nombreux traitements de texte ;
* des fonctions d'administration.

TRADUCTIX INC.
ATAO

ATAO est un progiciel léger, performant et adaptable qui offre des fonctions intégrées de dépouillement, de gestion terminologique et de prétraduction automatique.

* Le dépouillement consiste à rechercher dans un texte les chaînes de mots ayant

un intérêt terminologique ou phraséologique, *et éventuellement d'opérer un recoupement avec un fichier terminologique existant.* Un concordancier facilite la consultation des contextes.

* La prétraduction automatique (PTA) consiste à produire une traduction partielle (prétraduction) du texte. Le traducteur transforme ensuite la version prétraduite en sa traduction finale, *avec l'aide d'un pavé de commandes spéciales.*

JOHN CHANDIOUX EXPERTS-CONSEILS INC.
METEO, LE GÉNÉRAL TAO, TA-AU-CN

METEO® : Système utilisé depuis 17 ans par Environnement Canada pour la traduction automatique de l'anglais vers le français, et inversement, des prévisions météorologiques destinées au grand public.

LE GÉNÉRAL TAO : Système se rapprochant de la génération automatique de traductions préenregistrées, utilisé par la Confédération Vie pour produire des contrats d'assurance et divers autres documents de nature répétitive. Langues : anglais-français, français-anglais.

TA-AU-CN : Système développé pour la traduction automatique des descriptions des articles contenus dans la base de données centrale de pièces détachées et fournitures du Canadien National.

OBSERVATOIRE QUÉBÉCOIS DES INDUSTRIES DE LA LANGUE (OQIL)

Le Québec a mis sur pied un observatoire des industries de la langue au cours de l'année 1989.

L'**Observatoire québécois de la langue (OQIL)** a pour principales fonctions d'effectuer une veille technologique, de favoriser la concertation et la collaboration entre les industriels, les chercheurs, les universitaires et l'Administration et enfin d'orienter les choix nationaux en matière d'industries de la langue.

Les mandats de l'**OQIL** consistent à établir un bilan des industries de la langue et à élaborer une réflexion stratégique sur ces dernières de façon à permettre de mieux formuler les orientations à privilégier en matière de recherche et de développement. L'**OQIL** doit également devenir un véritable lieu de rencontre de tous les intervenants afin d'assurer un développement cohérent des industries de la langue tant sur le plan national que sur le plan international tout en assurant la diffusion de l'information.

L'**OQIL** ainsi que les observatoires canadien, français, suisse, wallon et africain des industries de la langue font partie d'un réseau plus vaste, le **Réseau international des observatoires francophones des industries de la langue (Riofil)**, dont le secrétariat général est assumé par le Québec.

SOCIÉTÉ CANADIENNE DE TRADUCTION ASSISTÉE (SOCATRA)
XLT

XLT est un système de TAO opérationnel et performant ; une réponse concrète à plusieurs de vos interrogations.

Index

Achevé d'imprimer sur
les presses de la SIEL (Beyrouth)
en avril 1995

La collection **Universités francophones** créée en 1988 à l'initiative de l'UREF, propose des ouvrages de référence, des manuels spécialisés et des actes de colloques scientifiques aux étudiants des 2e et 3e cycles universitaires ainsi qu'aux chercheurs francophones et se compose de titres originaux paraissant régulièrement.

Leurs auteurs appartiennent conjointement aux pays du Sud et du Nord et rendent compte des résultats des recherches et des études récentes entreprises en français à travers le monde. Ils permettent à cette collection pluridisciplinaire de couvrir progressivement l'ensemble des enseignements universitaires en français.

Enfin, la vente des ouvrages à un prix préférentiel destinés aux pays du Sud tient compte des exigences économiques nationales et assure une diffusion adaptée aux pays francophones.

Ainsi la collection **Universités francophones** constitue une bibliothèque de référence comprenant des ouvrages universitaires répondant aux besoins des étudiants de langue française.

Prix : 140 FF • Prix préférentiel UREF (Afrique, Asie, Amérique du Sud, Moyen-Orient) : 60 FF

59.46.61.1

9 782909 611099

ISBN 2-909611-09-4