HU

Linguistique

Linguistique et Traitements Automatiques des Langues

Catherine Fuchs
Laurence Danlos
Anne Lacheret-Dujour
Daniel Luzzati
Bernard Victorri

HACHETTE
Supérieur

HACHETTE UNIVERSITÉ
LANGUE LINGUISTIQUE COMMUNICATION

Collection dirigée par Bernard Quemada et François Rastier

Linguistique et traitements automatiques des langues

par

Catherine Fuchs

avec la collaboration de
Anne Lacheret-Dujour
et de
Bernard Victorri

ainsi que le concours de
Laurence Danlos
et de
Daniel Luzzati

HACHETTE
Supérieur

TITRES DÉJÀ PARUS DANS LA COLLECTION

P. Charaudeau : *Langage et discours – Éléments de sémiolinguistique*
J.-C. Coquet : *Sémiotique – L'École de Paris*
J. Courtés : *Sémantique de l'énoncé : applications pratiques*
J. Fontanille : *Les espaces subjectifs – Introduction à la sémiotique de l'observateur*
A.-J. Greimas et E. Landowski : *Introduction à l'analyse du discours en sciences sociales*
P. Lerat : *Sémantique descriptive*
M. Mathien : *Le système médiatique : le journal dans son environnement*
M. Meyer : *Logique, langage et argumentation*
R. Moreau : *Introduction à la théorie des langages* (épuisé)
Ch. Muller : *Initiation aux méthodes de la statistique linguistique*
Ch. Muller : *Principes et méthodes de la statistique lexicale* (épuisé)
J. Pinchon : *Morphosyntaxe du français – Étude de cas*
F. Rastier : *Sens et textualité*
A. Silbermann : *Communication de masse – Éléments de sociologie empirique*

NOUVELLE SÉRIE *LANGUE*

R.-L. Wagner et J. Pinchon : *La grammaire du français classique et moderne* (nouvelle édition)

NOUVELLE SÉRIE *LINGUISTIQUE*

J.-L. Chiss, J. Filliolet, D. Maingueneau : *Linguistique française – Syntaxe, discours, poétique* (nouvelle édition)
J. Courtés : *Analyse sémiotique du discours – De l'énoncé à l'énonciation*
J. Courtés : *Sémiotique narrative et discursive*
C. Fuchs et P. Le Goffic : *Les linguistiques contemporaines – Repères théoriques* (nouvelle édition)
C. Fuchs : *Linguistique et traitements automatiques des langues*
A.-J. Greimas et J. Courtés : *Sémiotique – Dictionnaire raisonné de la théorie du langage* (nouvelle édition)
A. Jaubert : *La lecture pragmatique*
D. Maingueneau : *L'analyse du discours – Introduction aux lectures de l'archive* (nouvelle édition)
D. Maingueneau : *L'Énonciation en linguistique française*
B. Pottier : *Théorie et analyse en linguistique* (nouvelle édition)

NOUVELLE SÉRIE *LINGUISTIQUE*

R. Escarpit : *L'information et la communication – Théorie générale* (nouvelle édition)
M. Mathien : *Les journalistes et le système médiatique*
R. Vion : *La communication verbale*

ISBN 2-01-016908-5

SOMMAIRE

AVANT-PROPOS

Ce livre est un ouvrage d'initiation, destiné à tous ceux (étudiants, enseignants, ...) qui désirent se familiariser avec les problématiques des traitements automatiques des langues. Il s'adresse à des non-spécialistes, et ne nécessite aucune connaissance préalable du domaine, en particulier aucune connaissance en informatique.

L'originalité de cet ouvrage réside dans le parti pris que nous avons retenu d'adopter un point de vue **linguistique** sur le domaine, et de présenter aussi bien les traitements de la langue **orale** que ceux de la langue **écrite**. Les linguistes trouveront une présentation des types d'approches de la langue que pratiquent les traitements automatiques, des difficultés de formalisation auxquelles ils se heurtent, ainsi que des modalités concrètes d'utilisation et de mise en oeuvre informatique des connaissances linguistiques ; ils découvriront les compromis nécessaires à l'élaboration de systèmes opérationnels, mais aussi les voies nouvelles de collaboration inter-disciplinaire qui s'offrent à eux. Les informaticiens, de leur côté, trouveront un exposé des problématiques linguistiques sous-jacentes aux différents secteurs des traitements automatiques et toucheront du doigt, nous l'espérons, la complexité et la spécificité des faits de langue, ainsi que la nécessité de procéder à des descriptions linguistiques fines des phénomènes à traiter.

L'ouvrage se compose de deux grandes parties. La **première partie** est consacrée aux **niveaux de traitement de la langue** orale et écrite : phonétique et phonologie, prosodie, morphologie, syntaxe, sémantique – étant entendu que ces types de connaissances linguistiques ne sont pas nécessairement traités dans autant de modules indépendants ; pour l'écrit, cette présentation est effectuée dans la perspective de l'**analyse** (les problèmes spécifiques de la génération faisant l'objet d'un développement dans la seconde partie). La **seconde partie** porte sur les **domaines** des traitements automatiques des langues : traitement de la parole, traduction automatique, compréhension automatique, génération automatique de textes, dialogue homme-machine ; là encore, tout en évoquant les types de systèmes réalisés ou en cours de réalisation, nous avons privilégié les problématiques linguistiques (les produits commercialisés dans le cadre des "industries de la langue" sont présentés dans l'Introduction).

Chaque chapitre est suivi d'une section intitulée **"Repères Bibliographiques"**, qui propose au lecteur des conseils de lecture ultérieure

au triple plan des descriptions et théories linguistiques, des formalismes et des implémentations informatiques.

Une **Introduction** assez développée permettra au lecteur de trouver quelques repères initiaux pour aborder la lecture de l'ouvrage ; nous lui conseillons (s'il en a le courage) de la reprendre en fin de lecture, afin d'opérer sa propre synthèse théorique.

Catherine FUCHS

INTRODUCTION

LES TRAITEMENTS AUTOMATIQUES DES LANGUES

Pour introduire aux problématiques qui seront exposées dans le cours du présent ouvrage, le lecteur trouvera ici quelques pistes de réflexion destinées à baliser le terrain de ce que l'on est convenu d'appeler les "traitements automatiques des langues" : nous tenterons de voir ce que recouvre cette expression (§ 1.), puis de cerner la diversité des applications réunies sous le terme "industries de la langue" (§ 2.), et enfin de situer les enjeux théoriques pour la recherche fondamentale (§ 3.).

1. Qu'entend-on par "traitements automatiques des langues"?

L'objectif des traitements automatiques des langues est la conception de logiciels (programmes) capables de **traiter** de façon **automatique** des données linguistiques, c'est-à-dire des données exprimées dans une **langue** (dite "naturelle"). Mais que recouvre au juste chacun des trois termes contenus dans cette expression "traitements automatiques des langues"? Nous considérerons tout d'abord le terme "langue", qui désigne l'objet de ces traitements automatiques, puis celui d'"automatique", enfin celui de "traitement".

1.1. "Langue"

C'est à dessein que nous avons retenu l'expression "traitement automatique ***des langues***", plutôt que celle – largement répandue – de "traitement

automatique *du langage*". Cette dernière dénomination, pour courante qu'elle soit, n'en est pas moins impropre. Elle décalque en effet l'expression anglaise "(natural) language processing" : or en anglais, le même mot "language" signifie à la fois "langage" et "langue" ; le français au contraire, distingue ces deux termes et, comme nous allons le voir, ce sont les ***langues*** (et non pas le *langage*) qui constituent le support des données linguistiques que les traitements automatiques ont pour objectif de manipuler.

L'objet des traitements automatiques qui nous occupent ici, ce sont des **données linguistiques**. Ces données peuvent être de différents types : il peut s'agir de textes écrits (au sens classique du terme, c'est-à-dire des suites de phrases constituant un tout informatif cohérent, comme par exemple une notice d'entretien, ou la description d'un appareil), ou bien de dialogues (écrits ou oraux), ou encore d'unités linguistiques de taille inférieure à ce que l'on appelle habituellement des textes (comme par exemple des phrases, des énoncés, des groupes de mots ou simplement des mots isolés). Par abus de langage, nous emploierons désormais le terme de "**texte**" pour désigner l'un quelconque de ces types de données linguistiques qui sont l'objet des traitements automatiques.

Qui dit "traitement" dit manipulation d'un **objet d'entrée**, aboutissant à la modification de cet objet en un nouvel objet (**objet de sortie**). Selon la nature de l'application, les traitements que nous considérons ici peuvent viser à agir sur un texte pré-existant en le transformant (par exemple pour le corriger, ou en extraire l'information, ou le condenser, ou le traduire, etc.), ou bien à créer un texte (en le construisant à partir d'informations données). Dans le premier cas (transformation d'un texte pré-existant), l'objet d'entrée est donc un texte (T) — dont la nature est matériellement très différente selon qu'il s'agit d'un texte oral ou d'un texte écrit — et l'objet de sortie peut être une représentation de ce texte (R(T)) utilisable par exemple pour la recherche ultérieure d'informations dans une base de données, ou pour faire exécuter des ordres par un robot, etc. ; ce peut être aussi un nouveau texte (T') qui paraphrase, résume ou traduit dans une autre langue le texte T après en avoir construit une représentation dans une étape intermédiaire. Dans le second cas (création d'un texte), l'objet de sortie est un texte (T) et l'objet d'entrée une représentation (R) de ce texte à venir, c'est-à-dire une représentation du contenu que l'on veut voir exprimé par le texte — ce contenu provenant lui-même d'une source extérieure, par exemple un autre logiciel ayant recherché l'information dans une base de données, dans une banque d'images, etc. Le lecteur trouvera dans les chapitres 6 à 10 qui composent la seconde partie de cet ouvrage une présentation détaillée des problèmes spécifiques posés par ces différents types d'applications (selon qu'il s'agit de traiter un texte oral ou écrit, qu'il s'agit de traduire ou de comprendre un texte pré-existant, de générer un texte nouveau, ou de construire des réponses aux questions d'un utilisateur). Dans tous les cas, on voit que les textes sont abordés comme des

objets que l'on peut **manipuler**, c'est-à-dire tester, modifier, agencer, construire ou reconstruire.

Pour pouvoir traiter un objet, il faut en connaître les principes de constitution interne, c'est-à-dire être capable de le **décrire** de façon **opératoire**. Un texte doit donc pouvoir être décrit comme un ensemble de formes (ou de correspondances entre formes et sens) régi par des règles explicitables : les **règles de la langue** ; les traitements automatiques se trouvent ainsi entraînés, de fait, à décrire (tout ou partie) de la langue qui constitue le support des textes à traiter. Soit dit au passage, aborder la langue comme un objet d'étude scientifique ne va pas de soi et exige un effort de décentration : la langue nous est en effet consubstantielle, dans la mesure où nous en avons inconsciemment intériorisé les règles en tant que sujets parlants (nous y reviendrons au § 3.1.).

Ainsi donc, c'est bien aux **langues** que les traitements automatiques se trouvent confrontés. Ceci nous conduit à rappeler quelques points de terminologie, au sujet de la différence, en français, entre "le langage" (humain), "les langages" (par exemple de programmation) et "les langues" (humaines).

Le langage désigne une faculté, proprement humaine (même si l'on parle métaphoriquement du "langage" des abeilles ou de l'apprentissage du "langage" par des chimpanzés, pour désigner des systèmes — assez limités au demeurant — de communication) ; le langage humain est, rappelons-le, la faculté de communiquer à l'aide d'un système de signes **doublement articulé** (la "première articulation" étant le niveau du morphème, c'est-à-dire de la plus petite unité significative ; la "deuxième articulation" étant le niveau du phonème, c'est-à-dire de la plus petite unité distinctive) — toutes les langues humaines sont ainsi structurées. Cette faculté de langage, innée, se développe avec l'enfant, lors de l'acquisition de la langue maternelle (puis des langues secondes), et peut se trouver (partiellement ou totalement) détruite accidentellement : acquisition et pathologie du langage font l'objet de recherches dans le domaine psycho- et neuro-linguistique.

Il est évident que la faculté de langage par elle-même ne peut constituer l'objet des traitements automatiques : cette faculté n'est pas observable directement, elle s'actualise nécessairement à travers l'utilisation d'une **langue** donnée ; la linguistique elle-même, "science du langage", n'appréhende celui-ci qu'au travers de la diversité des langues. Encore une fois, l'objet linguistique que l'on cherche à traiter automatiquement n'est pas le langage, mais des textes exprimés dans une langue particulière.

Ajoutons toutefois que les recherches en traitement automatique des langues qui se situent dans une perspective dite d'"intelligence artificielle" visant d'une manière ou d'une autre à simuler le comportement humain face à des textes (cf. *infra*, § 3.3.) se trouvent amenées à prendre en compte, par-delà les règles spécifiques au système de telle ou telle langue donnée, des phénomènes plus généraux, qui relèvent davantage du **langage** que d'une langue particulière : il s'agit, d'une part de phénomènes **trans-linguistiques**

(comme par exemple l'ambiguïté, la paraphrase, la métaphore et autres figures, l'ellipse, etc.), et d'autre part de mécanismes **langagiers** (comme par exemple les mécanismes de raisonnement et d'interprétation permettant, lors de la compréhension, d'inférer une signification plus complète à partir du contenu informatif littéral décodé).

On ne confondra pas "le langage" (au singulier) ainsi défini, avec **"les" langages** (au pluriel), tels qu'ils sont étudiés en informatique par la "théorie des langages" (branche de l'informatique qui s'intéresse aux propriétés mathématiques des différents types de langages de programmation). C'est précisément afin d'opposer ces langages artificiels que sont les langages de programmation, au langage humain, que l'on a forgé, pour désigner ce dernier, le néologisme "langage naturel". Les langages de programmation sont des **codes** (comme le sont aussi le code de la route ou le morse), forgés de toutes pièces par l'homme, fixes et immuables, totalement explicites ; les langues humaines au contraire ne sont pas des codes : leur origine est lointaine, elles évoluent constamment, et comportent une part importante d'implicite, des marges de jeu, des ambiguïtés (nous retrouverons ce point plus bas au § 3.2.).

De même que l'on parle de "langage naturel", on parle également de "langues **naturelles**" : ce terme ne devrait être utilisé, en toute rigueur, que pour opposer des langues comme le français, le chinois, le swahili, etc., aux "langues artificielles" que sont les langues que l'homme a tenté de construire (comme par exemple l'espéranto).

Enfin il convient de souligner que, pour des raisons économiques et politiques, seules les grandes langues vernaculaires des pays développés (et, au tout premier chef, l'anglais) se trouvent faire l'objet, de façon tant soit peu systématique, de recherches et d'applications en matière de traitements automatiques (cf. *infra*, § 2. : la notion d'"industrialisation" d'une langue). En dehors de toute considération géopolitique, disons qu'au plan théorique, cette situation de fait est regrettable pour le progrès des connaissances. En effet, la grande tradition de la linguistique générale avait, depuis près d'un siècle, amassé des trésors d'observations sur les langues les plus diverses du globe, montrant d'une part qu'il n'existe pas de "petite" langue du point de vue de l'organisation cognitive sous-jacente, et d'autre part que les systèmes linguistiques diffèrent considérablement d'une langue à l'autre. Les traitements automatiques ne risquent-ils pas de nous faire oublier cet acquis et de nous ramener à un ethnocentrisme linguistique appauvrissant et réducteur ? N'oublions pas, par ailleurs, que pour un certain nombre de langues, la perspective d'un traitement automatique entraîne des problèmes spécifiques : sans parler des langues à tradition orale ne possédant pas de système d'écriture, songeons aux problèmes de traitement de l'écriture pour des langues comme l'arabe ou comme le japonais (avec son double système de caractères *kana* et *kanji*).

1.2. "Automatique"

Est dit "automatique" un traitement qui opère par des moyens "mécaniques" (du grec *mêkhanê "machine"*), par opposition à un traitement manuel ou instrumental opéré par l'humain. La machine est ici l'**ordinateur**, c'est-à-dire une machine conçue pour effectuer des **calculs** : cela suppose donc que l'on soit capable de ramener les manipulations sur les données linguistiques à des calculs. Qui dit "traitement automatique" dit "suite d'actions (ici de calculs) à faire effectuer par la machine dans un certain ordre chronologique", c'est-à-dire un **programme.**

L'automatisation du traitement peut être **totale** ou **partielle** : on distingue de ce point de vue les traitements entièrement **automatiques** (effectués en totalité par l'ordinateur, sans intervention de l'humain dans l'exécution de la tâche) et les traitements **assistés par ordinateur** (effectués pour partie par l'ordinateur et pour partie par l'humain : soit que l'humain intervienne pour préparer ou terminer une tâche intermédiaire automatisée, soit qu'il recoure à l'ordinateur comme à un auxiliaire à certains moments précis d'une tâche manuelle nécessitant gain de temps et / ou précision de calcul). C'est ainsi par exemple que l'on parle, selon les cas, de "traduction automatique" ou de "traduction assistée par ordinateur" (comme on parle d'"enseignement assisté par ordinateur", de "publication assistée par ordinateur", etc.) : cf. *infra*, § 2.3.

Traiter un objet linguistique de façon **automatique** implique un certain nombre de **contraintes** dans la description même de cet objet : il faut pouvoir arriver à formuler de façon totalement **explicite** et **cohérente** des ensembles de **règles** caractérisant le fonctionnement du texte. Il faut donc savoir observer l'objet pour en dégager des régularités généralisables, et savoir exprimer ces régularités en éliminant le flou, l'implicite, le non-dit, les évidences allant de soi, et en vérifiant la cohérence de la formulation, qui garantit l'objectivité et donc la reproductibilité de la démarche : à cet égard, l'objectif d'un traitement automatique impose aux descriptions linguistiques des exigences de rigueur, de systématicité et de cohérence tout à fait salutaires.

Peut-être enfin n'est-il pas inutile de rappeler que la notion de traitement "automatique" des langues appelle une certaine **démystification**. Comme toute entreprise d'automatisation de tâches "intelligentes" jusqu'alors effectuées par l'humain, les traitements automatiques des langues donnent lieu à des réactions ambivalentes de fascination / répulsion : la machine pourra-t-elle réellement simuler le comportement langagier des humains ? n'y a-t-il pas dans le langage, activité humaine intelligente par excellence, une part de subjectivité irréductible, impossible à reproduire de façon automatique ? de même n'y a-t-il pas des types d'objets linguistiques irréductibles à un traitement automatique (textes littéraires, poétiques, faisant l'objet de jugements de valeur esthétiques) ? la langue n'est-elle pas surtout faite d'"exceptions" inexplicables, d'expressions "idiomatiques" ? etc. Soyons réalistes : l'ordina-

teur ne fait que ce qu'on lui dit de faire, ne sait (comprend) que ce qu'on sait (comprend) et qu'on est capable de lui expliciter ; sa seule supériorité réside dans sa puissance de calcul et (éventuellement) dans la taille de sa mémoire. Retournons donc les questions vers nous : pour l'instant nous ne savons traiter (et encore...) que des textes à visée informative, correspondant à des univers référentiels précis et restreints — et ceci aux niveaux les plus "bas" du traitement (très grossièrement dit : plus près des formes que du sens). Les progrès en matière de traitements automatiques des langues viendront de nos capacités à décrire les mécanismes de langue de façon plus fine, et à appréhender des régularités cachées derrière d'apparentes "exceptions".

1.3. "Traitement"

"Traiter" évoque l'idée d'agir sur un objet, en lui faisant subir un certain travail : en le manipulant, en le transformant, voire en le créant. Les objets pouvant donner lieu à des traitements automatiques sont de nature très diverse : langage, images, sons, etc.

Pour pouvoir traiter un objet, il faut disposer d'**outils** et de **techniques** de traitement. En matière de traitements automatiques des langues, ceux-ci sont de trois ordres : linguistiques, formels et informatiques. En effet, si l'observation et la description des mécanismes de langue supposent la maîtrise de connaissances et de techniques **linguistiques** précises, la description d'une langue en vue d'un traitement automatique nécessite en outre le recours à des procédures formelles réglées permettant d'exprimer les connaissances linguistiques dans des **formalismes** susceptibles d'être implémentés en machine ; de plus, pour effectuer, sur un texte donné, un traitement automatique, la machine doit appliquer des techniques **informatiques** lui permettant de mettre les connaissances qu'elle possède au service de l'application désirée. En toute rigueur théorique, il convient donc de distinguer la description des connaissances (notamment linguistiques), l'expression de ces connaissances dans un formalisme, et l'élaboration de techniques et de stratégies informatiques de traitement effectif. Dans la pratique toutefois, comme on le verra plus loin (cf. *infra*, § 3.1.), les traitements automatiques ne mettent pas toujours en oeuvre ces trois types d'outils (il arrive par exemple que seuls soient utilisés des outils informatiques), et quand ils y recourent, ils peuvent leur donner des poids respectifs très variables.

Ajoutons pour terminer que l'on ne saurait parler "du" traitement, mais "des" traitements automatiques des langues : il existe en effet une **diversité** de traitements, qui sont fonction à la fois des objectifs visés (la nature du texte et le type de manipulation effectué dépendent du domaine d'application), des théories linguistiques utilisées, des modélisations et des moyens informatiques mis en oeuvre.

1.4. En résumé

Nous retiendrons de ce qui vient d'être dit, que **les traitements automatiques des langues** ont pour objet des **données linguistiques (textes)** exprimées dans une **langue ("naturelle")**, et que pour pouvoir traiter automatiquement ces données, il faut être capable d'expliciter les **règles de la langue**, de les représenter dans des **formalismes** opératoires et calculables et de les implémenter à l'aide de **programmes**.

Le lecteur non-spécialiste trouvera à la fin de la présente introduction (au § 1. des repères bibliographiques) quelques pistes de lecture lui permettant, s'il le désire, de cerner ou d'approfondir les problématiques générales des traitements automatiques des langues.

2. Enjeux sociaux : les industries de la langue

Le besoin de logiciels de traitement automatique de textes en langue "naturelle" s'est progressivement fait jour durant ces dernières décennies, et devient à l'heure actuelle de plus en plus pressant, sous l'effet conjugué du développement de l'informatique (notamment de la micro-informatique), et de la nécessité de traiter rapidement une masse toujours croissante d'informations écrites ou parlées.

L'**informatique** envahit, on le sait, tous les secteurs de la société, non seulement dans les domaines socio-économiques, mais aussi dans la vie quotidienne des individus : plutôt que de devoir s'initier longuement et péniblement à une diversité de langages-machine, le non-informaticien souhaiterait pouvoir utiliser sa propre langue pour dialoguer avec un ordinateur "convivial".

Par ailleurs, nos sociétés contemporaines se trouvent confrontées à des quantités chaque jour plus importantes d'**informations** véhiculées en langue naturelle : d'où le besoin de méthodes rapides et efficaces permettant d'une part de composer et de structurer les documents à produire, d'autre part, de classer, d'archiver, d'analyser, d'interroger et de traduire les documents déjà produits.

2.1. Enjeux économiques

La conception d'ordinateurs conviviaux dialoguant en langue naturelle d'une part, la gestion automatique de l'information en langue naturelle d'autre part constituent des enjeux économiques considérables : l'élaboration de logiciels de traitement automatique des langues est donc appelée à connaître un essor considérable dans les années à venir. Un responsable de la MIDIST (Mission interministérielle de l'information scientifique et tech-

nique), J-F. Dégremont, écrivait déjà, en 1984 (p. 5) : "La langue est un élément primordial de la production industrielle : une centrale nucléaire représente 100 000 pages de documentation technique. Une automobile, un central téléphonique, un ordinateur, sont autant de produits qu'il n'est pas question de produire, réparer, exporter sans faire un important effort de rédaction et de traduction technique dont le coût représente un pourcentage important du prix total du produit (...)". Les traitements automatiques des langues conduisent ainsi à des gains de productivité dans de nombreux secteurs industriels (au premier chef dans le domaine de l'informatique et de l'électronique, mais aussi dans bien d'autres secteurs de l'économie).

2.2. Enjeux politiques, idéologiques et culturels

Enjeux économiques de première importance, les traitements automatiques des langues constituent également un enjeu politique, dont les États ont progressivement pris conscience. L'invasion des marchés par des logiciels réalisés en anglais apparaît en effet comme de nature à déstabiliser les langues nationales : pensons aux premiers traitements de texte, conçus pour l'anglais, qui ne prenaient pas en compte les signes diacritiques du français (tels l'accent, la cédille ou le tréma), et ne respectaient pas les normes typographiques françaises pour la coupure des mots en fin de ligne, ou encore aux programmes de jeux ou d'enseignement assisté par ordinateur qui, à travers la langue anglaise, imposent à l'utilisateur tout le système culturel américain.

Désireux de relever ce défi lancé à toutes les langues minoritaires, l'ensemble des pays d'Europe s'est mobilisé pour apporter un soutien concerté à la recherche en matière de traitement automatique des langues : "Les langues qui ne s'industrialiseront pas cesseront, à un terme plus ou moins bref, d'être véhiculaires, d'être des langues de civilisation" déclarait en 1985 le secrétaire général du Conseil de l'Europe. Dans cet esprit, B. Cassen rédigeait en décembre de la même année, à la demande du Ministère de la Recherche et de la Technologie, un rapport intitulé "Les industries de la langue : un grand enjeu culturel, scientifique et technologique pour la France", et le Conseil de l'Europe organisait en février 1986 à Tours un colloque international sur les industries de la langue, réunissant quelque 500 spécialistes du domaine, qui débouchait sur le principe d'une "maquette de mise à plat (c'est-à-dire de description systématique et exhaustive) de la langue". Dans le même ordre d'idées, le contrat européen ESPRIT II "Polyglot", démarré en 1989, vise à développer une architecture commune intégrant les modules nécessaires à la synthèse vocale à partir d'un texte écrit, pour six langues européennes (hollandais, anglais, français, allemand, grec et italien). On pourrait également citer le lancement en 1982 du projet EUROTRA de traduction automatique entre les neuf langues de la Communauté Economique Européenne (cf. *infra*, chapitre 7), ou encore le programme GENELEX (français à l'origine, puis associant des partenaires italiens et espagnols dans un projet EUREKA depuis

1990), visant à l'élaboration d'un modèle standard évolutif pour la conception de dictionnaires électroniques normalisés.

Il est clair que cette question de l'industrialisation de la langue est d'autant plus sensible dans les pays où une langue subit une situation de domination : on s'explique aisément que le Québec soit l'un de nos partenaires les plus actifs en matière de francophonie et d'industrialisation du français.

2.3. Les industries de la langue

D'outil de communication entre les hommes, la langue est ainsi devenue un **produit** faisant l'objet de développements industriels. D'où l'apparition de ce que l'on est convenu d'appeler, depuis le début des années 80, les "industries de la langue" (pour une présentation synthétique, voir l'Introduction de l'ouvrage de R. Carré *& al* 1991, pp. 7-25 ; d'autres références complémentaires sur les industries de la langue sont également données au § 2 des repères bibliographiques qui figurent à la fin de la présente introduction).

On regroupe sous le terme "industries de la langue" un ensemble très diversifié de travaux qui contribuent à la construction ou à l'amélioration de logiciels **commercialisés** manipulant des données linguistiques, de façon **opérationnelle**, dans des domaines **limités** : ces logiciels, qui s'appuient sur des bases de connaissances linguistiques nommées "linguiciels" (ou "linguisticiels"), abordent la langue surtout au plan des formes (ainsi les traitements de texte et correcteurs orthographiques), et ne travaillent qu'à l'intérieur d'univers référentiels restreints (comme par exemple les systèmes d'interrogations de données factuelles ou bibliographiques en langue naturelle, les lexiques et dictionnaires de spécialité, ou encore les programmes d'enseignement assisté par ordinateur).

De nombreux laboratoires de recherche, tant publics que privés, sont engagés dans l'élaboration de tels produits commercialisés ; de même un certain nombre de programmes et d'organismes nationaux ou européens financent des travaux dans ce domaine des industries de la langue (pour un aperçu, voir l'Annexe intitulée "Qui fait quoi ?", pp. 285-293 de l'ouvrage de R. Carré *& al.* 1991). Signalons également la création, dans plusieurs pays, d'"Observatoires des Industries de la Langue" (en France, l'Observatoire Français des Industries de la Langue publie une revue intitulée *La Tribune des Industries de la Langue*, et organise régulièrement des congrès sur ce thème).

Il n'est sans doute pas inutile de souligner que le secteur des industries de la langue offre des **débouchés** professionnels, non seulement pour la conception, la réalisation, l'exploitation et la maintenance de produits, mais aussi au plan de ce que l'on appelle l'"ingéniérie linguistique" (ou encore le "génie linguistique") : on désigne ainsi l'ensemble des activités de conseil et d'expertise auprès d'utilisateurs ayant des besoins dans le domaine des indus-

tries de la langue, ainsi que les activités de planification en vue de l'industrialisation des langues (mise sur pieds de programmes de recherche, conception d'appels d'offres, suivi d'une politique linguistique, etc.).

A cet égard, les propos tenus en 1984 (p. 5) par J-F. Dégremont sont toujours d'actualité : "De même que l'automatisation des travaux de production repose sur l'utilisation de sciences telles la chimie, la physique, la mécanique, l'informatique, de même l'automatisation des travaux de service repose pour partie sur la **linguistique** et ses différentes branches (...) La profession d'**ingénieur-linguiste** apparaît dans les sociétés de service et de conseil en informatique, et la demande dans cette spécialité est loin d'être satisfaite tant qualitativement que quantitativement".

Pour revenir aux **produits** offerts par les industries de la langue, précisons qu'ils concernent des secteurs très variés. Une première grande distinction peut être faite entre d'une part les réalisations portant sur l'oral (la parole) et d'autre part les réalisations portant sur l'écrit.

Les réalisations concernant l'**oral** (voir *infra*, chapitre 6) ont trait d'une part à la **reconnaissance** de la parole, d'autre part à la **synthèse** vocale. Les systèmes actuellement commercialisés sont assez limités ; ainsi en reconnaissance de la parole, des conditions fortes sont imposées aux systèmes : monolocuteur (ou multilocuteur sur un nombre de mots limité), apprentissage, nécessité de parler doucement, sans bruit de fond, etc. Toutefois des applications existent déjà dans plusieurs domaines : aide à l'éducation de la parole pour enfants sourds grâce à un système de visualisation des sons produits par l'enfant comparés avec la courbe obtenue par l'orthophoniste, installations expérimentales de commande de l'environnement pour grands handicapés, machines à dictée automatique, commande vocale de robots, etc. Les réalisations dans ce domaine font appel à la physique du signal, à l'acoustique, à la phonétique ; l'amélioration des systèmes existants passera par des recherches plus fondamentales s'appuyant aussi sur d'autres types de connaissances linguistiques (prosodiques, syntaxiques et même sémantiques).

Les réalisations concernant l'**écrit** ont trait principalement aux traitements de textes, aux systèmes d'extraction d'informations et aux systèmes de traduction .

Les **traitements de texte** offrent, en premier lieu, les ressources d'une super-machine à écrire dotée d'une mémoire et de diverses fonctions (tri, mise en page, impression, utilisation de plusieurs polices de caractère, mise en mémoire pour corrections, déplacement ou ajout de simples mots ou de passages entiers, recherche de mots, fabrication automatique d'index et de bibliographies, etc.). Certains logiciels traitent en outre de graphismes particuliers : formules mathématiques, alphabets non latins, braille, etc. Par ailleurs, les traitements de textes évolués proposent certaines fonctions plus complexes : correction orthographique (mais les problèmes difficiles, comme par exemple ceux des accords, qui supposent un calcul syntaxique, ne sont en

général pas traités), éléments de correction de phrases grammaticalement mal construites, et de correction stylistique (détection des répétitions ou des formulations lourdes, proposition de synonymes, calcul d'un indice de lisibilité, etc.) ; on s'achemine donc vers des traitements plus sophistiqués, qui par-delà la simple prise en compte des formes, tendent à travailler aussi au niveau du contenu (cf. *infra*, § 2.4.).

L'**extraction d'informations** consiste à alimenter des systèmes informatiques, en particulier des **bases de données**, à partir d'informations contenues dans des textes écrits. Ces systèmes peuvent être très frustres et se contenter de repérer dans des documents l'apparition de mots ou de groupes de mots qui servent ensuite à **indexer** ces documents dans une base documentaire. Mais ils peuvent aussi aller jusqu'à extraire des informations précises, calculées à partir du texte, qui alimentent une **base de données factuelles**, procédant ainsi à une véritable compréhension du texte (nous en verrons des exemples au chapitre 8 *infra*). On peut également placer ici l'activité de **résumé automatique**, qui en est encore à ses débuts (cf. D. Le Roux 1990 et 1992), et qui consiste d'abord à extraire l'information "essentielle" d'un texte d'entrée, pour construire à partir de cette information, un nouveau texte de sortie, condensé : on se rapproche ainsi de la problématique de la traduction ou de la paraphrase.

La **traduction automatique**, quant à elle (voir *infra*, chapitre 7), a été historiquement l'un des secteurs-pionniers du traitement automatique de l'écrit, puisque les premiers essais remontent à la fin de la dernière guerre. On parle de traduction automatique à propos des systèmes qui effectuent eux-mêmes l'ensemble des opérations de traduction, sans intervention de l'humain ; on oppose ces systèmes aux systèmes de traduction assistée par ordinateur, où l'humain intervient pour préparer le texte d'entrée, réviser le texte de sortie, ou aider à la traduction proprement dite. Les systèmes de traduction comportent trois phases de traitement linguistique : l'analyse du texte d'entrée en langue-source, la génération du texte de sortie en langue-cible, et entre les deux, le transfert de la représentation du texte d'entrée en une représentation du texte de sortie. La plupart des systèmes commercialisés sont limités à des domaines bien précis (avionique, traitement des textiles, bulletins météo, ...), et traduisent des textes scientifiques ou techniques de structure relativement simple et régulière, parfois au prix d'une normalisation du texte d'entrée, réduit à un sous-ensemble des structures de la langue nommé "sous-langage". De nombreux produits annexes (outils informatisés d'aide à la traduction humaine, dits "postes de travail du traducteur") sont développés dans le champ des industries de la langue.

L'**articulation** des traitements de l'**écrit** et de l'**oral** constitue, à terme, l'un des objectifs des traitements automatiques des langues. Elle est déjà effective dans certaines réalisations qui relèvent du domaine de la communication homme-machine en langue "naturelle" : les systèmes de **dialogue**

homme-machine (voir *infra*, chapitre 10) sont des interfaces en langue naturelle, permettant à l'humain de converser avec l'ordinateur. Très utiles dans un certain nombre de secteurs, comme par exemple celui de l'interrogation de bases de données, ces systèmes comportent deux moments de traitement linguistique : l'analyse de la question de l'utilisateur, et la génération d'une réponse par la machine (ou d'une question relançant le dialogue, en cas d'échec de la compréhension). De tels systèmes sont d'ores et déjà opérationnels sur des dialogues finalisés dans des univers limités (horaires de train, services, ...).

Comme on l'aura compris à travers cette brève esquisse des réalisations en matière d'industries de la langue, les systèmes opérationnels à l'heure actuelle sont **limités**, tant au plan de leurs domaines d'application (il s'agit toujours d'univers fermés très restreints facilement modélisables) qu'au plan des performances linguistiques (ils ne traitent pas la langue dans toute sa généralité, mais seulement certains aspects de la langue, certaines formes ou certaines structures canoniques relativement simples à décrire et à formaliser). Pour répondre aux besoins, et pouvoir concevoir des systèmes plus perfectionnés et plus généraux capables de dépasser les limites des systèmes actuels, il apparaît nécessaire de se tourner vers des recherches plus théoriques ou fondamentales.

2.4. Des applications aux recherches fondamentales

En amont des industries de la langue à vocation appliquée, et susceptible de nourrir celles-ci de retombées substantielles à plus ou moins long terme, se situe en effet tout un ensemble de recherches plus fondamentales (menées en particulier dans le cadre de laboratoires universitaires et du CNRS) : ces recherches s'attachent à couvrir la langue de la façon la plus large possible, et à en modéliser le fonctionnement à des niveaux plus "profonds" (syntaxique, sémantique, pragmatique,...). Les systèmes ainsi élaborés en sont encore bien souvent au stade expérimental de la maquette : rares sont les prototypes fonctionnant en grandeur réelle et *a fortiori* les produits finis commercialisables.

De telles recherches fondamentales s'avèrent indispensables si l'on veut, à terme, améliorer les systèmes existants, et en concevoir de nouveaux, plus performants. Considérons, à titre d'exemple, les systèmes de **traitement de texte** : comme on l'a vu plus haut, les systèmes commercialisés, élaborés dans le cadre des industries de la langue, manipulent les textes à un niveau de **formes** assez élémentaire (caractères et mots) ; mais on s'achemine progressivement vers des systèmes manipulant des formes plus complexes (à savoir des structures syntaxiques) et même travaillant au plan du **contenu** (en intégrant des considérations sémantiques, et en s'orientant vers la compréhension et la génération de textes).

De même en ce qui concerne plus spécifiquement les **correcteurs ortho-**

graphiques : il est clair qu'ils visent désormais à dépasser le seul niveau de la consultation d'unités simples (mots) codées dans un dictionnaire, et à intégrer des règles syntaxiques (voire même sémantiques et pragmatiques) pour être en mesure de corriger notamment les fautes d'accord (qui sont, comme l'on sait, parmi les plus fréquentes et les plus gênantes pour les sujets) ; divers prototypes ont été élaborés dans cette perspective (cf. *infra*, chapitre 4), et certains systèmes effectuant des analyses syntaxiques partielles commencent à apparaître sur le marché.

De même encore, en matière de conception de systèmes de **traduction**, on assiste à l'heure actuelle, dans les milieux de la recherche, à l'élaboration de prototypes qui visent à effectuer une analyse beaucoup plus "profonde" du texte d'entrée (intégrant davantage de sémantique, et même de pragmatique) avant de procéder au transfert du texte dans la langue-cible, afin d'essayer d'améliorer la qualité des traductions (cf. *infra*, chapitre 7).

De même enfin, dans le domaine du **dialogue** homme-machine, l'amélioration des systèmes nécessite une modélisation plus poussée des interactions verbales, intégrant des considérations pragmatiques (cf. *infra*, chapitres 8 et 10).

Pour toutes ces applications, l'élaboration de systèmes plus performants passe donc par le détour de recherches fondamentales en matière notamment de **compréhension de texte** (voir *infra*, chapitre 8) et de **génération de texte** (voir *infra*, chapitre 9). Dans ces deux perspectives, le traitement de la langue porte non seulement sur les formes, mais aussi sur le contenu ; il doit mettre en oeuvre des **connaissances linguistiques** très complètes (relevant des niveaux de la morphologie, de la syntaxe, de la sémantique et de la pragmatique), ainsi que des connaissances d'univers. La plupart des systèmes actuels sont encore expérimentaux et font appel à des techniques d'intelligence artificielle.

Il va sans dire que de telles recherches revêtent nécessairement un caractère pluri-disciplinaire, et doivent associer étroitement linguistes et informaticiens. Ceci nous conduit à dire quelques mots sur les enjeux théoriques des traitements automatiques des langues.

3. Enjeux théoriques : l'ordinateur face à la langue

Le linguiste sait comment fonctionne la langue, l'informaticien sait comment marche l'ordinateur : dès lors, apprendre à l'ordinateur les règles de la langue pour lui faire traiter des données linguistiques peut, à première vue, paraître un jeu d'enfant. Pourtant, l'interdisciplinarité linguistique / informatique s'avère, dans la pratique, extrêmement difficile à réaliser : force est donc de questionner les apparentes évidences qui viennent d'être énoncées.

3.1. Linguistique et informatique : un dialogue difficile

La difficulté tient à deux ordres de raisons théoriques (sans préjudice des facteurs socio-économiques qui ont entraîné, comme chacun sait, un déséquilibre certain entre les deux disciplines informatique et linguistique ... au profit de la première des deux!) :

- en premier lieu, il existe, entre l'humain concepteur de systèmes de traitement et l'objet linguistique à modéliser en vue du traitement, un lien étroit d'ordre tout à la fois cognitif et affectif : un effort de **décentration** est nécessaire au sujet parlant (linguiste ou informaticien) pour aborder la langue, dont il a inconsciemment intériorisé les règles, comme un objet de connaissance et d'expérimentation scientifique ;

- en second lieu, le rapport de l'informatique à l'objet langue est d'une autre nature que le rapport habituel de l'informatique à un simple domaine d'application (par exemple lorsque l'informaticien construit, avec l'aide du spécialiste de la discipline concernée, un système expert en chimie ou en médecine) : il existe en effet entre le fonctionnement de l'objet à traiter (la langue) et le fonctionnement de la machine (les langages de programmation) des relations **analogiques**, à tel point que l'on a pu être tenté d'assimiler purement et simplement la langue à un langage de programmation.

Cette situation très particulière conduit l'informaticien et le linguiste à adopter, spontanément, face à la langue, des positions diamétralement opposées. L'**informaticien**, s'appuyant sur sa compétence de sujet parlant (éventuellement assortie de quelques souvenirs de grammaire scolaire), se croit "naturellement" capable de décrire la langue, qu'il tend à aborder de façon réductrice à l'image d'un langage formel, méconnaissant ainsi la spécificité et la complexité de l'objet à décrire (ajoutons toutefois qu'à cet égard, le linguiste "entiché" de formalismes peut pratiquer, avec tout le zèle du néophyte, un réductionnisme plus féroce encore...). S'il est vrai qu'une telle attitude tend plutôt à régresser (à mesure que les concepteurs de systèmes s'attaquent à des phénomènes linguistiques de plus en plus difficiles, et que la pluridisciplinarité gagne du terrain dans les milieux de la recherche), elle reste néanmoins suffisamment répandue pour que l'on y insiste tant soit peu : ce n'est pas parce que l'on est doué, comme tout être humain, de la faculté de langage, et que l'on a appris une ou plusieurs langues, que l'on est pour autant capable de décrire objectivement et adéquatement les mécanismes de langue. Que l'on nous autorise, sur ce point, une comparaison : le téléphone est un objet que des millions d'usagers savent utiliser ; pour autant, peu d'entre eux sauraient décrire de façon scientifique la manière dont il fonctionne (chacun s'accordant à reconnaître que c'est là affaire de spécialistes). En matière de langue aussi, la description relève de spécialistes : les linguistes. La comparaison s'arrête toutefois là, car l'objet langue est incomparablement plus complexe et mal connu qu'un objet technologique fabriqué et entièrement contrôlé par l'homme.

Le **linguiste** de son côté, tout en évitant de tomber dans le travers d'une attitude de fusion subjective avec l'objet (qui, dans sa version extrême "littéraire", n'autoriserait pour toute description que l'infinité des commentaires incontrôlables, des reformulations explicatives, ou des paraphrases herméneutiques), a une conscience aiguë de la complexité et de l'hétérogénéité des facteurs constitutifs de la langue, et est habitué à rencontrer des phénomènes apparemment rebelles à toute mise en équations. Cette fréquentation de la langue tend à le rendre méfiant à l'égard des formalismes et des modélisations (qu'il juge facilement réducteurs), et le conduit par ailleurs à effectuer préférentiellement des descriptions fines et parcellaires de micro-dömaines de langue (en s'attaquant de façon privilégiée aux phénomènes les plus difficiles à expliquer), plutôt qu'à chercher à "couvrir" superficiellement sinon la totalité de la langue, du moins un maximum de structures canoniques simples.

D'où, en retour, une insatisfaction de l'**informaticien**, constatant que les règles de fonctionnement de la langue sont loin de faire l'objet de descriptions unanimes, complètes et opératoires, de la part des spécialistes que sont les linguistes : chaque école linguistique avance en effet sa propre conception de la langue, choisit son terrain privilégié de faits à décrire, et opte pour un arsenal théorique particulier ; l'informaticien se trouve ainsi confronté à une myriade de descriptions rivales incomparables entre elles (tant au plan de la terminologie que des concepts théoriques) et qui, pour systématiques et formelles qu'elles puissent être, n'en restent pas moins partielles, difficilement généralisables, et malaisément implémentables. Il est clair que cette situation reflète moins une quelconque insuffisance de la discipline linguistique que la complexité de l'objet à décrire ; pour autant, on comprend le désarroi et la défiance de l'informaticien à l'égard de la linguistique.

Le lecteur estimera peut-être que nous avons grossi le trait. Il n'en reste pas moins vrai d'une part que les intérêts et les points de vue de l'informaticien et du linguiste en matière de description de la langue sont par nature différents, et d'autre part que l'état de nos connaissances sur la langue est encore trop lacunaire pour permettre de couvrir l'ensemble de la langue en la décrivant de façon homogène à l'aide d'un métalangage facilement implémentable — *a fortiori* lorsque l'on quitte le plan des formes (morphologie et syntaxe) qui depuis longtemps fait l'objet d'études à vocation formelle, pour aborder la sémantique et la pragmatique linguistiques (terrains d'investigation plus récents).

Si, dans son principe même, le domaine des traitements automatiques des langues appelle une collaboration à égalité entre les deux disciplines linguistique et informatique, force est de reconnaître que dans la pratique une telle collaboration (qui, par-delà un apprentissage mutuel des terminologies et des perspectives, passe par la construction progressive d'une problématique commune) n'est pas chose fréquente : attestée dans certains secteurs de recherche fondamentale, elle l'est beaucoup moins en matière de développement de

systèmes, et encore moins de fabrication de produits, où l'informatique est très nettement dominante.

La différence de dominance possible entre les deux disciplines est censée se refléter, en français, dans les deux dénominations respectives d'"informatique linguistique" et de "linguistique informatique". L'**informatique linguistique** est conçue dans son principe comme une branche de l'informatique, qui ne recourt qu'aux méthodes et outils de l'informatique, et dont le domaine d'application se trouve être des données linguistiques. Dans un sens étroit, l'expression "informatique linguistique" renvoie à l'utilisation de logiciels (comme par exemple des logiciels effectuant des statistiques) pour opérer des calculs sur les mots ou suites de mots contenus dans un texte (calcul de cooccurrence, etc.) : toute une série de travaux (qui ne seront pas abordés dans le présent ouvrage) s'inscrivent dans ce courant, comme par exemple les travaux dans le domaine de la lexicométrie. Prise dans un sens large, l'expression en vient à désigner l'ensemble des traitements automatiques de données linguistiques.

La **linguistique informatique** (anglais "computational linguistics"), quant à elle, est – dans son principe du moins – une branche de la linguistique formelle, qui recourt à l'ordinateur comme outil de validation d'hypothèses théoriques sur le fonctionnement de la langue : elle a vocation à se servir des techniques informatiques pour expérimenter sur des données linguistiques, tester des systèmes de règles, étudier la pertinence linguistique de certaines modélisations, voire simuler certains comportements langagiers. Si la linguistique informatique est effectivement conçue et pratiquée de la sorte dans certaines équipes de recherche fondamentale, précisons toutefois c'est une autre conception qui prévaut dans les milieux de la recherche appliquée : y sont dits relever de la "linguistique informatique" tous les travaux en traitement automatique des langues qui, d'une manière ou d'une autre, s'appuient sur des éléments d'analyse linguistique (ce qui inclut bon nombre de réalisations de type "industries de la langue", et ne préjuge en rien de l'importance de la contribution linguistique).

En définitive, le problème de fond soulevé par ces dénominations concerne le mode **d'articulation** entre les **données** linguistiques d'entrée et leur **traitement** par la machine : faut-il une théorisation linguistique des données ? faut-il une étape intermédiaire de formalisation entre cette théorisation et le traitement informatique ? A ces deux questions, l'informatique linguistique tend à répondre par la négative, et la linguistique informatique par l'affirmative (ainsi que nous l'avons fait, pour notre part, dès le § 1.1. *supra*). Même lorsqu'est admise la nécessité de recourir aux trois types d'outils (linguistiques, formels et informatiques), leur mode d'articulation peut être très différent (cf. J.P. Desclés 1986 p. 34) ; schématiquement, on peut distinguer trois grands types d'approches de ce problème :

- certains traitements partent directement des implémentations **informatiques** et ne recourent qu'à un minimum de théorisation linguistique, sans

autre modélisation que les représentations informatiques elles-mêmes : de tels systèmes ne sont, par nature, pas généralisables (c'est le cas, par exemple, des premiers systèmes documentaires d'interrogation par mots-clés, et aussi de certains analyseurs très rudimentaires) ; nous dirons de cette approche qu'elle est "pilotée par l'informatique".

- au contraire des précédents, certains traitements partent de théories **linguistiques** donnant lieu à des types de formalismes : la question est alors de faire le lien avec le traitement informatique et d'arriver à implémenter ces formalismes, sans préjudice par ailleurs de la question de la pertinence linguistique des formalismes retenus (c'est notamment le cas des traitements syntaxiques recourant à des grammaires formelles — cf. *infra*, chapitre 4 § 3. — ainsi que des sémantiques formelles fondées sur des "logiques intensionnelles"— cf. *infra*, chapitre 5 § 2.2.) ; nous dirons de cette approche qu'elle est "pilotée par la linguistique (formelle)".

- enfin un troisième type de traitement prend pour point de départ un **formalisme** emprunté à la **logique** (le calcul des prédicats) et se tourne simultanément vers l'informatique (utilisation de ce formalisme à des fins de programmation — ce que l'on appelle la "programmation logique" avec le langage PROLOG) et vers les données linguistiques (utilisation du même formalisme à des fins de représentation de ces données) : la question est évidemment celle de la double adéquation du formalisme, et en particulier de son adéquation pour rendre compte des données linguistiques, tant syntaxiques (voir par exemple les "definite clause grammars" : cf. *infra*, chapitre 4 § 4.4) que sémantiques (sur les limites de la logique classique comme formalisme de représentation de la sémantique linguistique, cf. *infra*, chapitre 5 § 2.2.) ; nous dirons de cette approche qu'elle est "pilotée par la logique".

La nécessité de recourir à un **formalisme** de représentation (de nature logique ou mathématique) comme **intermédiaire** entre les données linguistiques à traiter et le traitement informatique proprement dit s'éclairera, nous l'espérons, à travers les brefs rappels qui suivent – destinés, bien évidemment, au non-informaticien (ce dernier trouvera quelques suggestions de lectures introductives sur l'informatique au § 3. des repères bibliographiques figurant à la fin de la présente introduction).

3.2. Ordinateurs, langages de programmation et langues naturelles

L'**informatique** a pour objet le traitement automatique de l'information, à l'aide d'un ordinateur ; celui-ci se définit donc comme une machine programmée pour traiter automatiquement de l'information. L'organe de traitement de l'ordinateur est le **processeur** (anglais "processor") – on parle de "micro-processeur" si cet organe tient dans un circuit intégré – qui, avec la mémoire, constitue le "cerveau" de la machine dit "unité centrale" ; cette

unité centrale est appelée ainsi pour la distinguer des éléments "périphériques" que sont d'une part le clavier, destiné à entrer les données à traiter ainsi que les programmes de traitement, et d'autre part l'écran et l'imprimante, destinés à sortir les résultats du traitement.

Les informations contenues dans la mémoire de l'ordinateur et qu'il doit traiter sont des caractères codés avec un **code numérique binaire** (rappelons que "numérique" — on dit aussi "digital" — s'oppose à "analogique", comme le discontinu s'oppose au continu, et que "binaire" signifie que le nombre de valeurs est de deux).

Les ordres de traitement sont des procédures (**algorithmes**) permettant à la machine d'accéder aux données (informations) et de les manipuler. Les **données** sont elles-mêmes représentées dans l'ordinateur de façon **structurée** (par exemple sous forme de listes) : ce sont ces "représentations internes" des données dans l'ordinateur que celui-ci est capable de traiter.

Les ordres de traitement ne sont pas donnés directement à l'ordinateur en langage-machine, mais transmis au processeur par l'intermédiaire d'un **langage de programmation** : face à un problème à traiter, l'informaticien, après une phase d'**analyse** des actions à faire effectuer, procède à un travail de **programmation**, c'est-à-dire d'écriture de ces actions dans un langage de programmation donné, puis il entre ce programme dans la machine à l'aide de l'"éditeur" (traitement de texte). L'ordinateur traduit le langage de programmation en langage-machine soit au fur et à mesure de l'exécution (à l'aide d'un programme appelé **interpréteur**), soit avant l'exécution (à l'aide d'un programme appelé **compilateur**).

Il existe une grande diversité de **langages de programmation**. Pour le traitement automatique de données linguistiques, les deux langages les plus couramment utilisés sont des langages développés dans le cadre de l'intelligence artificielle, à savoir LISP (qui présente l'avantage de permettre une manipulation facile des structures de listes) et PROLOG (dû à A. Colmerauer, cet outil de "programmation logique", qui permet d'intégrer des mécanismes de déduction, a été retenu pour la "cinquième génération" d'ordinateurs). (Le lecteur désireux d'approfondir la question des méthodes de programmation pour le traitement automatique des langues, en particulier à l'aide de PROLOG, pourra se reporter aux ouvrages mentionnés au § 4. des repères bibliographiques figurant à la fin de la présente introduction).

Les données linguistiques ne peuvent pas être transcrites directement en représentations internes de l'ordinateur ; c'est pourquoi il est nécessaire de passer par l'intermédiaire d'une **représentation formelle** : le recours à un formalisme permet d'associer aux données linguistiques d'entrée (textes) une représentation formelle, qui elle-même doit ensuite être transformée en représentation interne pour l'ordinateur.

On appellera **langages formels** aussi bien les langages de programmation que les formalismes de représentation intermédiaire dont il vient d'être ques-

tion (par exemple langages logiques). Tout comme le langage-machine, ces langages formels sont des **codes** (qu'ils soient destinés à un interpréteur ou à un compilateur). Comme nous l'avons dit plus haut (cf. *supra*, § 1.3.), les langues naturelles, elles, n'ont pas les propriétés d'un code ; elles évoluent dans le temps, elles comportent nécessairement de l'implicite, et elles ne connaissent pas de correspondance bi-univoque entre forme et sens : c'est précisément cette non-biunivocité constitutive entre le plan des signifiants (marqueurs et structures de marqueurs) et le plan des signifiés (valeurs sémantiques) — source des phénomènes d'ambiguïté, de polysémie, de synonymie et de paraphrase — qui donne aux langues cette marge de jeu, cette labilité leur permettant d'être des instruments de communication (et pas seulement des moyens de consigner de l'information).

Cette différence tout à fait essentielle entre une langue et un code explique les difficultés que rencontrent les traitements automatiques dès lors qu'ils s'efforcent de traiter, dans toute leur complexité, de "vrais" textes en langue naturelle : même lorsqu'il est à vocation informative, un texte ne se ramène jamais à un simple codage d'informations ; le calcul de sa sémantique est plus complexe que l'interprétation de symboles dans un univers donné (cf. *infra*, chapitre 5), et l'élaboration d'un système de compréhension de textes passe nécessairement, par-delà une analyse sémantique des phrases et des relations inter-phrastiques, par la prise en compte de paramètres contextuels et situationnels (cf. *infra*, chapitre 8), *a fortiori* s'il s'agit de dialogues (cf. *infra*, chapitre 10).

C'est pourtant vers les langages formels que l'on s'est tourné pour aborder le traitement automatique des langues, assimilant "traitement automatique de données linguistiques" à "traitement automatique de l'information" et "langue" à "langage formel". En matière de langages formels, il est classique de se référer à la tripartition syntaxe / sémantique / pragmatique : la **syntaxe** concerne la combinatoire des symboles permettant de construire des expressions bien formées, la **sémantique** concerne l'interprétation des symboles dans un univers donné, la **pragmatique** concerne l'utilisation des symboles par des usagers. Cette tripartition a été réimportée pour décrire les langues naturelles : syntaxe, sémantique et pragmatique étant abordées comme autant de niveaux de description autonomes. Pourtant, la pertinence d'une telle approche peut être mise en question : la définition des niveaux ne peut se superposer à celle que l'on connaît pour les langages formels, et leur autonomie est problématique.

La **syntaxe** linguistique a bien trait à la combinatoire des unités sur la chaîne syntagmatique, mais les règles de cette combinatoire ne permettent pas toujours d'opposer deux sous-ensembles totalement disjoints d'expressions (bien formées / mal formées) ; ces règles ne sont d'ailleurs pas aussi bien connues et univoquement explicitables que celles d'un langage formel. La **sémantique** linguistique, quant à elle, ne saurait se ramener à un

ensemble de règles d'interprétation des symboles dans un univers donné : le sens est à distinguer de la référence, comme des valeurs de vérité de la logique ; de plus, comme nous l'avons dit plus haut, le rapport forme-sens est beaucoup plus complexe qu'un rapport bi-univoque symbole-interprétation. Enfin la **pragmatique** linguistique ne renvoie pas à des conditions d'utilisation extrinsèques au système : à la fonction informative se mêlent toujours d'autres fonctions (rhétoriques, argumentatives, etc.) qui "mettent en scène" les protagonistes de l'acte de communication, et le système même de la langue comporte ses propres règles de mise en fonctionnement (déictiques, personnes, temps, modalités, etc.).

De surcroît, l'**autonomie** de ces trois niveaux est loin d'aller de soi : la syntaxe dépend de la sémantique car la construction des structures syntaxiques oblige parfois à recourir à des considérations d'ordre sémantique (voire même pragmatique) — cf. *infra*, chapitre 4 ; à son tour, la sémantique dépend de la pragmatique, car dans la signification d'un énoncé sont indissociablement mêlés le sens prédicatif et les valeurs référentielles-énonciatives (sans parler des cas où la signification première d'un énoncé correspond à une valeur illocutoire et non au sens dit littéral) — cf. *infra*, chapitres 5 et 8.

Prenant progressivement conscience de ces difficultés, les traitements automatiques ont ainsi été amenés à moduler leur conception de départ, ce qui a eu pour conséquence une modification des **architectures de systèmes**. Dans un premier temps, les trois composantes syntaxique, sémantique et pragmatique ont été conçues comme trois niveaux (**modules**) de traitement de la langue totalement indépendants et strictement hiérarchisés ; d'où des architectures **modulaires** en couches, qui ont reçu une diversité de dénominations : structures "hiérarchiques", "séquentielles", "stratificationnelles", ou encore "en série". Dans un système de ce type (efficace techniquement, car simple à mettre en oeuvre), chaque module prend en entrée les données qui lui sont fournies par la sortie du module précédent, et ainsi de suite dans un ordre fixe qui interdit toute rétroaction d'un module sur un autre ; les inconvénients pour le traitement de la langue sont clairs, du fait de l'interdépendance de fait des niveaux dans la langue : en se privant des informations du module $n+1$ qui permettraient d'effectuer un choix face à une multiplicité de solutions concurrentes au niveau n, on risque l'explosion combinatoire (les traitements engendrant par eux-mêmes des séries d'ambiguïtés parasites : cf. *infra*, chapitres 3 à 5).

Pour tenter d'éviter ces difficultés et de construire des systèmes plus souples, d'autres architectures moins hiérarchisées ont été proposées, qui se fondent essentiellement sur des systèmes de coopération et d'échange entre modules (systèmes **multi-experts**), le contrôle des traitements pouvant devenir très sophistiqué, comme par exemple dans les systèmes à "tableau noir" (que nous aurons l'occasion d'évoquer à plusieurs reprises dans cet ouvrage).

Les systèmes à architecture modulaire en couches tendent donc à être abandonnés, tant en analyse qu'en génération — du moins dans l'élaboration

de prototypes expérimentaux, car les solutions alternatives proposées, pour séduisantes qu'elles soient théoriquement, se heurtent à des difficultés de mise en oeuvre, en particulier lorsqu'il s'agit de passer à des systèmes en grandeur réelle. En effet, il faut expliciter toutes les interactions entre les différents types d'informations pour pouvoir ordonner, à mesure des besoins, les interventions des différents experts ; ce qui suppose, au plan linguistique, que l'on sache évaluer le poids relatif des diverses règles (de tous niveaux) contribuant au calcul (d'un mot, d'une structure, d'un sens, etc.) : là encore, on est loin de savoir maîtriser, de façon effective, le jeu d'un grand nombre de paramètres.

L'évocation de ces architectures de systèmes plus souples, qui relèvent de ce que l'on est convenu d'appeller l'"intelligence artificielle", nous conduit à dire quelques mots, pour terminer, de cette branche de l'informatique et des travaux qui s'y mènent en vue du traitement automatique de textes en langue "naturelle".

3.3. Langage, cognition et intelligence artificielle

Il n'entre évidemment pas dans notre propos de faire une présentation d'ensemble du domaine de l'intelligence artificielle ; le lecteur souhaitant s'initier à ce domaine trouvera des suggestions de lecture au § 3. des repères bibliographiques figurant à la fin de la présente introduction. Nous voudrions simplement essayer de montrer en quoi la démarche de l'intelligence artificielle a pu renouveler les problématiques des traitements automatiques des langues, et évoquer les recherches pluri-disciplinaires (incluant la linguistique et l'intelligence artificielle) qui se développent à l'heure actuelle dans le champ plus large des "sciences de la cognition" (sur ces questions, on pourra consulter les ouvrages et articles mentionnés au § 5. de nos repères bibliographiques).

Il est difficile de donner de l'**intelligence artificielle** une définition simple et unique, le domaine qu'elle couvre étant complexe et controversé. Rappelons pour mémoire que le terme d'"intelligence artificielle" a été forgé en 1956 par des scientifiques réunis pour étudier la possibilité de réaliser des programmes d'ordinateur doués d'intelligence : l'objectif visé par l'intelligence artificielle est en effet d'essayer de doter l'ordinateur de capacités habituellement attribuées à l'intelligence humaine (comme par exemple la perception visuelle ou auditive, le raisonnement, la prise de décision, et...le langage). L'objectif est donc plus **ambitieux** que celui des traitements automatiques classiques : il ne s'agit pas seulement de traiter de l'information à partir de données linguistiques, mais d'arriver à faire effectuer par l'ordinateur sur des textes ce que l'humain lui-même est capable d'effectuer (par exemple les comprendre, raisonner sur eux, etc.). On voit pourquoi ce sont moins les langues que le **langage** (en tant que comportement humain intelligent) qui est au centre des recherches : au fond, il s'agit de construire une "machine douée de langage" en lui inculquant des comportements langagiers intelligents.

Quelle que soit la définition que l'on en donne, on s'accorde généralement à reconnaître que l'intelligence artificielle adopte dans sa démarche un certain nombre de caractéristiques **méthodologiques** qui la distinguent des approches informatiques classiques (cf. J-P. & M-C. Haton 1989. pp. 6-9) : elle traite des informations de nature essentiellement **symbolique** (concepts, règles, objets, faits) — même si, bien entendu, les procédures de traitement elles-mêmes restent numériques —, elle recourt à des méthodes **heuristiques** (c'est-à-dire à des moyens permettant de trouver une solution rapide à un moindre coût, mais dont les éléments sous-jacents ne sont pas totalement explicités, contrairement aux méthodes **algorithmiques** classiques), elle peut traiter des données **incomplètes** ou **inexactes**, voire conflictuelles (en effectuant par exemple des raisonnements approximatifs).

Mais surtout, l'une des principales caractéristiques de l'intelligence artificielle est la place centrale qu'elle accorde à la notion de **connaissances** : pour résoudre des problèmes dans un domaine donné, il faut se fonder sur une énorme quantité de connaissances relevant de ce domaine ; il faut donc savoir **représenter** ces connaissances (elles sont généralement codées sous une forme dite "**déclarative**", par opposition à l'approche "**procédurale**" de l'informatique classique) et savoir les gérer — d'où l'importance des systèmes à **bases de connaissances** (notamment des **systèmes experts**). Cette question de la représentation des connaissances (linguistiques et d'univers) et de la modélisation d'un domaine se retrouve par exemple pour l'élaboration de systèmes de compréhension de textes, comme nous le verrons au chapitre 8 (cf. *infra*).

Symptomatique à cet égard est le changement terminologique qui conduit à substituer aux "informations" (objet du traitement de l'informatique classique) des "connaissances" : c'est ainsi qu'en matière de traitement de données linguistiques, les "niveaux de description de la langue" (morphologie, syntaxe, sémantique, ...) deviennent des "types de connaissances linguistiques". Comme nous l'avons dit plus haut, ces différentes composantes linguistiques ne sont plus "encapsulées" dans des modules étanches, mais traitées par des "experts" (c'est-à-dire des gestionnaires intelligents de connaissances) qui coopèrent. Il s'ensuit une plus grande souplesse dans la mise en oeuvre des mécanismes d'analyse ou de génération linguistique, qui constitue une réponse adéquate au constat de l'**interaction** entre paramètres morphologiques, syntaxiques, lexicaux, sémantiques, etc. : à cet égard, la démarche rejoint celle des linguistes qui, depuis longtemps déjà, parlent de phénomènes "morpho-syntaxiques", "syntaxico-sémantiques", voire même "pragma-sémantiques", montrant que les catégories de langue sont interreliées et se conditionnent mutuellement.

Par ailleurs, un autre aspect de cette approche des "connaissances" par l'intelligence artificielle concerne la question de l'**ordination des règles**. Dans un premier temps, l'idée était de pouvoir mettre toutes les connaissances "à plat", sans se soucier de les ordonner ; très vite pourtant, est appa-

rue la nécessité d'établir des priorités et l'on a alors recouru à un certain nombre de techniques permettant de classer les règles par paquets, et d'associer à ces paquets de règles des niveaux de priorité (c'est ce que l'on trouve, en gros, dans les systèmes experts classiques). Puis sont apparues d'autres techniques, plus sophistiquées (comme par exemple les "tableaux noirs" dont il a déjà été question) visant à implémenter de façon déclarative des **règles de contrôle** intelligent sur les règles elles-mêmes (ces règles de contrôle faisant partie des "méta-règles" ou "méta-connaissances"). L'idée est la suivante : il s'agit de permettre d'énoncer sur des classes de règles des **priorités conditionnelles** (c'est-à-dire relatives et non pas absolues) de type : "la classe de règles A a priorité sur la classe de règles B dans telles conditions..." ; autrement dit, cela permet d'expliciter les raisons de la priorité de la classe A sur la classe B dans certains cas de figure, sans être obligé d'énoncer un ordre fixe constant. On voit l'intérêt de ce type d'approche pour la linguistique : si, sur un domaine de faits de langue donné, le linguiste se trouve généralement dans l'impossibilité d'ordonner de façon absolue les règles qu'il établit, en les prenant une par une, en revanche il peut arriver —- moyennant une étude extrêmement détaillée et précise du micro-domaine de langue considéré — à élucider les raisons pour lesquelle telle règle tend à s'appliquer avant telle autre, sous telle et telle condition.

Malgré les avantages théoriques qui viennent d'être évoqués, le recours "banalisé" à la notion de connaissances n'est pas exempt, du point de vue du linguiste, d'un certain risque de "confusion des genres" : paradoxalement, vouloir traiter ensemble tous les types de connaissances impliquées dans un comportement langagier intelligent, conduit tout à la fois à mieux reconnaître la spécificité du fonctionnement de la langue (par différence avec un langage formel) et à nier la spécificité des mécanismes linguistiques (dans leur articulation aux mécanismes conceptuels et logiques). Comme on le verra aux chapitres 5 et 8 (*infra*) à propos de sémantique et de pragmatique, l'autonomie relative des signifiés (par rapport aux concepts et aux représentations d'univers) et des opérations linguistiques (par rapport aux opérations cognitives) tend à être largement ignorée ; sur ce point, on trouvera de très pertinentes critiques dans les travaux de F. Rastier cités au § 5. de nos repères bibliographiques. Au total, l'objet "langue" risque donc de se trouver **dilué** dans ces approches du "langage" par l'intelligence artificielle.

Ceci nous conduit à notre dernier point, celui de la **pluri-disciplinarité** des recherches sur le langage, susceptibles de réunir spécialistes d'intelligence artificielle et linguistes, sous la bannière plus large des "**sciences cognitives**" : on désigne par ce terme un ensemble de disciplines (de la neuro-physiologie et la biologie à la linguistique, la psychologie, etc.) concernées par l'étude de la "cognition" chez les êtres vivants (et plus particulièrement les humains) – étude qui recouvre notamment les champs suivants : perception, raisonnement, action, et **langage**. Dans ce domaine, les types de collaboration de la linguis-

tique avec d'autres disciplines sont multiples : par-delà les classiques que constituent les études menées depuis longtemps déjà en psycho-linguistique ou en neuro-linguistique, des collaborations plus larges s'esquissent avec la perspective de centres de recherches cognitives sur le langage (il en existe déjà par exemple aux Etats-Unis) réunissant notamment linguistes, psychologues, ergonomes, informaticiens, logiciens.

Dans un tel contexte, les travaux en traitement automatique sont susceptibles de prendre un nouveau tournant : la notion de **simulation** jouant alors un rôle-clé. A cet égard, deux conceptions de l'intelligence artificielle sont possibles : la conception "basse" (ou réaliste) qui prévaut actuellement, et qui consiste à tenter d'atteindre (d'"émuler") les mêmes résultats que l'humain face à une tâche donnée sans pour autant chercher à reproduire (à "simuler") les processus qui ont pu y conduire, et la conception "haute" (peut-être utopique), qui revient à essayer de reproduire les processus eux-mêmes, c'est-à-dire les opérations mentales sous-jacentes (postulées). En matière de comportement langagier (compréhension de textes, génération de textes, traduction, résumé, etc.), cette seconde approche passerait par la construction de systèmes automatiques **psychologiquement plausibles** : il s'agirait dans cette perspective de construire et de faire manipuler par l'ordinateur des représentations que l'on postule être en conformité avec un modèle des représentations mentales humaines ; si une telle perspective reste encore très largement du domaine de la science-fiction, quelques tentatives en ce sens commencent néanmoins à voir le jour (par exemple pour construire des modèles de génération de textes qui, par-delà la simple synthèse de formes d'expression à partir de représentations d'entrée, soient de véritables modèles de "production"). Dans de telles entreprises, la linguistique devrait bien évidemment avoir un rôle essentiel à jouer, en particulier pour rappeler la spécificité et la diversité des langues, dans des cercles où l'on ne parle que de "langage" (sous réserve toutefois qu'elle sache se faire entendre dans le concert des sciences cognitives où, à l'heure actuelle, du moins en France, sa présence est aussi faible numériquement que peu diversifiée au plan des approches théoriques).

Mais il y a loin de projets aussi ambitieux à leur mise en oeuvre : comme nous le verrons tout au long de cet ouvrage, les réalisations opérationnelles en matière de traitement automatique de données linguistiques sont modestes, et nécessitent beaucoup de réalisme dans la démarche (notamment dans l'évaluation des faits de langue traitables dans l'état actuel de nos savoir-faires), ainsi que beaucoup de temps et de travail pour la réalisation effective.

Osons, pour conclure cette introduction, poser la question qui nous a hantée au fil de l'élaboration de cet ouvrage : les traitements automatiques font-ils progresser nos connaissances sur la langue ? si oui, en quoi ? si non, pourquoi ?

Catherine FUCHS
(ELSAP-CNRS)

Repères bibliographiques

1. Introductions aux traitements automatiques des langues :

[Ouvrages et articles généraux, les uns de vulgarisation, les autres de synthèse :]

CARRE, R. *& al.* (1991) : *Langage humain et machine*, Paris, Presses du CNRS.

[Recueil de cinq contributions d'auteurs différents, consacrées respectivement aux industries de la langue (les enjeux, et les marchés), aux problématiques générales du traitement des langues naturelles, au traitement de l'écrit, au traitement de l'oral, et aux mécanismes de dialogue ; d'une lecture accessible aux non-spécialistes.]

COULON, D. & KAYSER, D. (1986) : Informatique et langage naturel : présentation générale des méthodes d'interprétation des textes écrits, *Technique et Science Informatiques*, 5 : 2, Paris, Gauthier-Villars, 103-128.

[Article de synthèse sur l'analyse automatique des textes écrits ; présente les grandes classes de modèles (syntaxiques, logiques, psycho-cognitifs), les divers types de modules de traitement, puis les principales applications (notamment pour l'interrogation de bases de données).]

DESCLES, J.P. (1986) : La linguistique informatique, *Le Courrier du CNRS*, 65, Paris, CNRS, 33-34.

[Très court article esquissant quelques notions théoriques constitutives de la linguistique informatique.]

FARGUES, J. & SABAH, G. (1992) : Intelligence artificielle et langage naturel, *La Recherche*, 245 : 23, Paris, 818-825.

[Article de vulgarisation sur les réalisations et les recherches en matière de traitement automatique dans une perspective d'intelligence artificielle.]

KAYSER, D. (1985) : Des machines qui comprennent notre langue, *La Recherche*, 16 : 170, Paris, 1198-1212.

[Article d'introduction aux traitements de l'écrit en vue de la compréhension de textes ; présente les différents niveaux de traitement, les types de modèles, et les problématiques sous-jacentes ; de lecture aisée pour le non-spécialiste, cet article bien documenté fait, en une dizaine de pages, un tour de la question.]

SABAH, G. (1988 / 1989) : *L'intelligence artificielle et le langage* ; vol. 1 : "Représentations des connaissances" / vol. 2 : "Processus de compréhension", Paris, Hermès.

[Ouvrage de synthèse extrêmement complet, le premier du genre en fran-

çais. Le volume I (352 pages) présente les théories linguistiques et les formalismes de représentation utilisés en traitement automatique ; le volume II (411 pages) traite des mises en oeuvre informatiques (dans une perspective d'intelligence artificielle). Consacré aux problématiques théoriques du traitement de textes écrits (compréhension, dialogue et génération) en intelligence artificielle, à l'exclusion des traitements de la parole, des produits en industries de la langue, et des systèmes de traduction automatique. Manuel de référence dans le domaine traité (plutôt destiné à la consultation qu'à la simple lecture), à recommander aux linguistes et aux informaticiens désireux de se spécialiser.]

SMITH, G. (ed.) (1991) : *Computers and human language*, San Mateo, Kaufmann.

[Ouvrage d'initiation très complet sur les problèmes, les méthodes et les outils du traitement automatique de la langue écrite et orale ; parcourt tous les niveaux de traitement, du phonème au discours. Chaque chapitre se termine par des exercices et des conseils de lecture. Peut constituer un guide de travail.]

TENNANT, H. (1981) : *Natural language processing : an introduction to an emerging technology*, Princeton, Petrocelli.

[Ouvrage introductif présentant les principaux systèmes américains de traitement automatique de l'écrit (anglais), élaborés avant le début des années 80 (premiers programmes sans analyse linguistique, analyseurs syntaxiques, analyseurs sémantiques, systèmes de représentation des connaissances d'univers, traitements du discours). Très descriptif, procède par "études de cas".]

WINOGRAD, T. (1984) : Les logiciels de traitement des langues naturelles, *Pour la Science*, nov. 84, Paris, 90-103.

[Article de vulgarisation introduisant aux logiciels de traitement automatique de textes ; présente successivement les problématiques de la traduction automatique, des traitements de textes avancés, des systèmes de dialogue, et des coordinateurs (auxiliaires de courrier électronique) ; insiste sur l'omniprésence de la question de l'ambiguïté.]

2. Industries de la langue :

ABBOU, A. *& al.* (eds.) (1987) : *Les industries de la langue : les applications industrielles du traitement de la langue par les machines* ; vol. I : "Analyse des concepts, des terrains, des technologies, des produits et des marchés en présence — financement et position de l'offre française face à la concurrence internationale" ; vol. II : "Introduction à la compréhension des technologies et à la connaissance des produits" ; Paris, DAICADIF.

[Le vol. I est une analyse de conjoncture, le vol. II une sélection d'articles de périodiques, souvent de vulgarisation, reproduits dans leur intégralité.]

ABBOU, A. (ed.) (1989) : *Répertoire des produits et services de traitement automatique de la langue française*, Paris, OFIL.

[Ce répertoire concerne les secteurs suivants : traitement de la parole ; traduction ; traitement de documents écrits ; exploration, analyse et génération de texte ; gestion de documentation assistée par ordinateur ; interfaces en langage naturel ; enseignement assisté par ordinateur.]

Actes du Colloque "Génie Linguistique 91" (Versailles, janvier 1991), EC2, 3 volumes.

[Voir en particulier le vol. 1 "Recherche et développement" ; les systèmes qui y sont présentés concernent les interfaces de consultation de bases de données, la reconnaissance de la parole, le dialogue vocal, la recherche documentaire, la traduction assistée par ordinateur, la correction syntaxique, les dictionnaires électroniques, et les applications des approches cognitives.]

Actes du Colloque "Traitement automatique de la langue et industries de l'information. Problématiques 1995" (Paris, Salon international des industries de la langue, novembre 1991).

[Recueil des interventions regroupées autour des thèmes suivants : la recherche industrielle, la formation en ingéniérie linguistique, la recherche universitaire, les relations entre concepteurs et distributeurs, les interfaces d'interrogation en langue natturelle, les correcteurs orthographiques et syntaxiques. Comporte également le catalogue des exposants, présentant divers produits.]

CARRE, R. *& al.* (1991) : *Langage humain et machine*, Paris, Presses du CNRS.

[Voir l'Introduction intitulée "Les industries de la langue : enjeux et stratégies, pp. 7-25, et l'Annexe intitulée "Qui fait quoi ?", pp. 285-294.]

DEGREMONT, J-F. (1984) : L'émergence d'une industrie de la langue, *BRISES*, 4, Paris, CDSH, 5-6.

LE ROUX, D. (1990) : Automatisation de l'activité résumante : vers une typologie, *Tribune des Industries de la Langue*, 3 : 8 / 10, Paris, OFIL, 2-5.

LE ROUX, D. (1992) : Vers l'automatisation du résumé de texte : quelques exemples de produits classés, *Tribune des Industries de la Langue*, 7-8, Paris, OFIL, 37-40.

Revue *La Tribune des Industries de la Langue*, Paris, Observatoire Français des Industries de la Langue.

[Revue trimestrielle française, qui paraît depuis le février 1990. Voir en particulier le n° 4 / 5 / 6 (nov. 90 / nov. 91) — numéro spécial intitulé "Ingéniérie linguistique : problématiques 1995".]

3. Informatique, intelligence artificielle :

- Quelques ouvrages de vulgarisation :

ANTORMARCHI, F. *& al.* (1986) : *Pense... machine : pour comprendre l'intelligence artificielle*, Paris, CESTA.

[Voir en particulier le chapitre 5 : "Langage naturel".]

BONNET, A. (1984) : *L'intelligence artificielle ; promesses et réalités*, Paris, Inter-éditions.

DREYFUS, H. (1984) : *Intelligence artificielle ; mythes et limites*, Paris, Flammarion.

GENTHON, PH. (1989) : *Dictionnaire de l'intelligence artificielle*, Paris, Hermès.

HATON, J-P. et HATON, M-C. (1989) : *L'intelligence artificielle*, Paris, P.U.F. (Que sais-je ?).

[Voir en particulier le ch. 5 : "Traitement du langage naturel.]

MATHELOT, M. (1969, rééd. 1991) : *L'informatique*, Paris, P.U.F. (Que Sais-je ?).

MORVAN, P. (1988) : *Dictionnaire de l'informatique ; concepts, matériels, langages*, Paris, Larousse.

RICH, E. (1983) : *Artificial intelligence*, New-York, McGraw-Hill ; trad. fr. (1987) : *Intelligence artificielle*, Paris, Masson.

SIMONS, G. (1985) : *Les ordinateurs de demain : la cinquième génération*, Paris, Masson.

WINOGRAD, T. & FLORES, F. (1986) : *Understanding computers and cognition*, Norwood, Ablex ; trad. fr. (1989) : *L'intelligence artificielle en question*, Paris, P.U.F.

[Voir en particulier la Partie II : "Calcul, pensée, langage".]

- Outre ces textes introductifs, on peut citer un ouvrage plus technique :

FARRENY, H. & GHALLAB, M. (1987, rééd.1990) : *Eléments d'intelligence artificielle*, Paris, Hermès.

4. Langages et méthodes de programmation pour le traitement automatique des langues :

COLMERAUER, A. (1984) : Prolog, langage de l'intelligence artificielle, *La Recherche*, 158 : 15, Paris, 1104-1114.

[Article de vulgarisation introduisant (sous la plume de son concepteur) au langage de programmation Prolog, l'un des plus utilisés pour le traitement automatique des langues.]

Outre cet article, on pourra consulter les ouvrages suivants, qui sont des manuels d'initiation au langage Prolog pour le traitement automatique des langues :

DAHL, V. & SAINT-DIZIER, P. (eds.) (1985) : *Natural language understanding and logic programming*, Amsterdam, North-Holland.

GARDENT, C. & BASCHUNG, K. (1993) : *Techniques d'analyse et de génération pour la langue naturelle*, Clermont-Ferrand, ADOSA.

GAL, A. *& al.* (1989) : *Prolog pour l'analyse automatique du langage naturel*, Paris, Eyrolles.

GAZDAR, G. & MELLISH, C. (1989) : *Natural language processing in Prolog : an introduction to computational linguistics*, Reading, Addison-Wesley.

MICHIELS, A. (1991) : *Traitement au langage naturel et Prolog*, Paris, Hermès.

Sur le langage de programmation LISP, on pourra consulter :

QUEINNEC, C. (1982) : *Langage d'un autre type : LISP*, Paris, Eyrolles.

5. Langage et cognition :

- Sciences cognitives :

BONNET, C. *& al.* (eds.) (1987) : *Psychologie, intelligence artificielle et automatique*, Bruxelles, Mardaga.

[Voir en particulier la Partie III : "Langage".]

IMBERT, M. *& al.* (eds.) (1987) : *Cognitive science in Europe*, Berlin, Springer.

STILLINGS, A. *& al.* (eds.) (1987) : *Cognitive science : an introduction*, Cambridge Mass., M.I.T. Press.

[Voir en particulier le ch. 6 : "Linguistics : the representation of language".]

VARELA, F. (1988) : *Connaître les sciences cognitives : tendances et perspectives*, Paris, Seuil.

- Langage, intelligence artificielle, cognition :

LENNY, J.F. (1989) : *Science cognitive et compréhension du langage*, Paris, P.U.F.

OSHERSON, D. & LASNIK, H. (eds.) (1990) : *Language, an invitation to cognitive science*, Cambridge Mass., M.I.T. Press.

RASTIER, F. (ed.) (1987) : "Sémantique et intelligence artificielle", *Langages*, 87, Paris, Larousse.

RASTIER, F. (ed.) (1989) : "Sciences du langage et recherches cognitives", *Histoire, Epistémologie, Langage*, 11 : 1, Paris.

RASTIER, F. (1991) : *Sémantique et recherches cognitives*, Paris, P.U.F.

SCHANK, R. & CHILDERS, P. (1984) : *The cognitive computer : on language, learning and artificial intelligence*, Reading Mass., Addison Wesley.

NB : Nous ne donnons aucune référence générale de linguistique dans cette introduction ; les suggestions de lecture en matière de linguistique seront indiquées chapitre par chapitre.

PREMIÈRE PARTIE

LES NIVEAUX DE TRAITEMENT DE LA LANGUE

1
PHONÉTIQUE ET PHONOLOGIE

Les connaissances phonétiques et phonologiques jouent un rôle fondamental tant en reconnaissance qu'en synthèse de la parole. En **synthèse à partir du texte**, une première étape que nous appellerons "**phonologique**" correspond à la phonétisation de n'importe quel texte écrit entré dans un système, pour fournir en sortie une chaîne de phonèmes. La seconde étape, de nature **phonétique**, consiste à associer des corrélats acoustiques à ces objets phoniques en tenant compte de leur assemblage, pour produire en finale un signal de parole continue. En **reconnaissance**, la démarche est inverse, puisque l'objet de départ est un continuum sonore, entité physique en soi, qu'il faut segmenter pour fournir en sortie des objets phoniques isolables les uns des autres. Cette étape **phonétique** n'est pas triviale, loin s'en faut ; en effet si la description d'un énoncé oral, telle qu'on la trouve dans tous les manuels d'initiation à la phonétique (cf. F. Carton 1974) et qui repose sur la notion de **phonème**, est pertinente du point de vue linguistique (le phonème est une unité fonctionnelle **distinctive**), elle n'en demeure pas moins relativement abstraite : l'image acoustique des phonèmes - objets théoriques construits par les linguistes - n'est pas identique en parole continue à celle des phonèmes isolés, le signal de parole est un continuum sonore complexe qui ne ressemble en rien à une simple concaténation d'unités discrètes. Le traitement **phonologique** consiste quant à lui à associer à une suite d'objets sonores donnés en entrée une suite de mots ; démarche également complexe puisque d'une part le mot n'a pas d'existence au niveau acoustique, d'autre part il faut tenir compte des phénomènes d'altérations contextuelles et individuelles des phonèmes.

1. Rappels théoriques

1.1. Phonétique et phonologie

Si la phonétique et la phonologie ont toutes deux pour objet les sons du langage, la démarche spécifique à chacune des deux disciplines justifie un rappel théorique. Dans une perspective **phonétique,** les sons du langage sont étudiés concrètement sous l'angle de leur production et / ou de leur transmission et de leur perception. La phonétique a par conséquent au moins deux champs d'investigation : le domaine acoustique étudie les caractéristiques acoustiques des sons produits et perçus, le domaine articulatoire ou physiologique est centré sur les études de l'appareil articulatoire et des mécanismes mis en oeuvre pour produire les sons du langage. Dans une perspective **phonologique**, et par-delà les oppositions d'école, le matériau phonique est appréhendé en vue de comprendre comment les sons du langage participent au fonctionnement de la langue et sont des unités fonctionnellement pertinentes au sein d'un système linguistique. Néanmoins, dans les faits cette séparation formelle entre phonétique et phonologie doit être nuancée : le phonéticien n'étudie jamais totalement les sons du langage indépendamment du code linguistique, et en phonologie c'est la même substance matérielle qui est analysée et interprétée ; les données descriptives habituellement exposées dans les manuels d'initiation à la phonétique doivent donc être connues par les phonologues. Cette interpénétration des disciplines est sans doute à l'origine de confusions et de désaccords terminologiques entre linguistes et chercheurs en traitement automatique de la parole.

1.2. Articulatoire et acoustique

Il n'est pas question ici de faire un exposé détaillé des mécanismes articulatoires mis en jeu dans la parole (sur ce point on pourra consulter Calliope 1989). Néanmoins, un bref rappel permettra de mieux saisir la complexité de la structure acoustique qui en résulte et la difficulté de corréler matière phonique et structure linguistique.

1.2.1. Complexité acoustique de la parole

Le signal de parole est la résultante physique de l'excitation, par une source sonore, du conduit vocal, associée dans certains cas à l'excitation du conduit nasal (pour la production des phonèmes nasals). Trois types de **sources** sonores sont à l'origine de la production des sons : la première est une source d'**impulsion périodique** constituée par l'ensemble poumons-cordes vocales, c'est elle qui permet de produire les **sons voisés** (les voyelles notamment). La seconde est une source de **bruit** qui produit un signal aléatoire apériodique caractéristique des **fricatives non voisées** ([s] [S] [f]). La dernière est

une source **impulsionnelle** liée à une suppression d'air à l'intérieur du conduit vocal et à une ouverture brusque de celui-ci. Elle est utilisée pour la production des **plosives sourdes** ([p] [t] [k]). Dans le premier cas (source d'impulsion périodique), les cordes vocales vibrent à une certaine fréquence, mesurée en hertz et déterminant la hauteur d'un son. Cette fréquence de vibration - ou **fondamental** (noté F0) - est liée aux caractéristiques intrinsèques des cordes vocales et à la pression subglottique correspondant au nombre de vibrations par seconde des cordes vocales. Les consonnes sonores quant à elles sont la résultante du couplage d'une impulsion périodique et d'une impulsion apériodique (source de bruit ou source impulsionnelle selon le type de consonne prononcée). Enfin, la production des consonnes nasales, également sonores, implique le couplage acoustique de deux cavités de résonance (pharyngo-buccale et nasale) : le conduit vocal est fermé pour que l'air s'écoule par les fosses nasales. Il en va de même pour la production des voyelles nasales, à une différence près : la cavité buccale reste ouverte dans sa partie antérieure.

Le signal émis au niveau des cordes vocales est décomposable en une somme d'oscillations sinusoïdales dont les fréquences (**harmoniques**) sont des multiples entiers du fondamental (**théorème de Fourier**). Ce signal est modifié en passant par le conduit vocal constitué par un ensemble de cavités (pharyngale, buccale, nasale) à forme et à volume variables. Leurs volumes respectifs et leur couplage acoustique sont déterminés par l'articulation et la position de la langue. Ces cavités jouent le rôle de caisses de résonance en renforçant le spectre des harmoniques dont la fréquence est proche de la leur. Les zones de fréquence ainsi renforcées sont appelées **formants**. Il s'agit ici ni plus ni moins des mêmes mécanismes physiques que ceux associés à la production de sons par des instruments à cordes tels que le violon ou la contrebasse. Précisons que si la forme et le volume du conduit vocal restent relativement fixes pour l'émission des voyelles, ils se modifient sensiblement pour la production des consonnes.

La résultante physique de ces mécanismes articulatoires est observable sur le signal, où l'on peut dégager quatre grands types de structures acoustiques :

- un signal régulier caractéristique des voyelles,
- un signal aléatoire bruité caractéristique des sons [s] et [S],
- des variations brutales associées à des bruits d'explosion qui correspondent à des ouvertures brusques du conduit vocal lors de la production des occlusives,
- un silence qu'il n'est pas toujours facile d'associer à une pause respiratoire ou à un élément constitutif d'une plosive.

En fonction de la source productrice de sons et des différents types de couplage acoustique, il est possible de classer les phonèmes selon leur **mode d'articulation**. Le **lieu d'articulation**, mesuré par le reserrement du conduit vocal provoqué par le dos de la langue, est également utilisé (cf. Figure I).

Mode d'articulation	Lieu d'articulation		
Consonnes	**Labiales**	**Dentales**	**Vélo-palatales**
Occlusives			
non voisées	p	t	k
voisées	b	d	g
nasales (voisées)	m	n	ɲ
Fricatives			
non voisées	f	s	ʃ
voisées	v	z	ʒ
Liquides		l	R
Glides	w	ɥ	j

Voyelles	**Antérieures**		**Postérieures**
	Non arrondies	**Arrondies**	
fermées	i	y	u
mi fermées	e	ø	o
mi ouvertes	ɛ	œ	ɔ
ouvertes	a	ɔ	
Nasales	**Antérieures**		**Postérieures**
fermées	ɛ̃		ɔ̃
ouvertes		ɑ̃	

Figure I
Les phonèmes du français

1.2.2. Variabilité de la parole

L'extrême variabilité du signal de parole est la source majeure de sa complexité et explique pour une grande part les difficultés rencontrées en traitement automatique de la parole. On peut distinguer trois sortes de **variantes** de prononciation : les variantes de co-articulation, les variantes combinatoires et les variantes libres.

Les variantes de **co-articulation** sont la conséquence de l'influence mutuelle des phonèmes. Elles s'expliquent par la nature même de la parole, qui est une structure continue. Les articulateurs utilisés pour produire tel ou tel son sont des systèmes mécaniques présentant une certaine inertie : leurs mouvements ne sont pas tous synchrones, il peut y avoir des retards, des anti-

cipations (mise en place anticipée d'un articulateur pour produire un son) et des réductions. Les réductions sont d'autant plus importantes que le débit d'élocution augmente : les gestes articulatoires sont moins précis et la cible articulatoire peut difficilement être atteinte. Les frontières des segments sont donc moins nettes.

Les variantes **combinatoires** (ou variantes "contextuelles") sont les prononciations allophoniques d'un seul et même son conditionnées par des contraintes contextuelles de nature linguistique : le contexte peut entraîner des phénomènes de **neutralisation**. Ainsi, l'opposition des phonèmes [e] [ɛ] est neutralisée dans le mot *serpent* où seule la prononciation ouverte est possible.

Enfin, les variantes **libres** constituent, indépendamment de contraintes linguistiques déterminées, l'ensemble des variantes liées à la constitution physiologique des locuteurs, à leur héritage sociolectal et à leurs origines géographiques. Ainsi, le conduit vocal d'un sujet féminin est en moyenne de 15% plus court qu'un conduit vocal masculin, le larynx est placé plus bas chez les hommes ; l'âge, le sexe, le milieu socio-professionnel ont des incidences sur le style vocal ; enfin un sujet du sud de la France ne parle certainement pas comme un québécois.

Si, méthodologiquement parlant, il est nécessaire de distinguer ces différentes variantes, dans la pratique, elles sont souvent intimement imbriquées. Ainsi, dans la séquence *prendre mon train,* l'aire de variation de la finale *'dre'* est assez souple, dans ce contexte plusieurs prononciations sont possibles : [prãdrɔ̃mɔ̃trɛ̃], [prãnrɔ̃mɔ̃trɛ̃] et [prãnmɔ̃trɛ̃]. Le choix entre ces trois prononciations pour un locuteur donné dépendra de ses habitudes langagières, de son débit d'élocution, enfin de son système phonatoire. Les variantes de prononciation peuvent également être déterminées par des contraintes syntaxiques et / ou prosodiques. Par exemple, la réalisation d'une liaison (facultative, interdite ou obligatoire) dépend de règles syntaxiques précises ; la durée d'un phonème peut être modulée par des variations prosodiques : une voyelle inaccentuée peut être réduite et, de ce fait, difficilement identifiable. Les sources de variabilité sont donc complexes et le lien sous-jacent entre ces différents niveaux de variation rend parfois difficile la formalisation à chaque niveau de l'analyse.

Ces variantes peuvent être regroupées dans deux grandes classes de **variabilité** : la variabilité **intra-locuteur** et la variabilité **inter-locuteur**. Nous définirons la première comme le champ de liberté phonétique d'un locuteur : toutes les manières possibles pour ce sujet de prononcer une séquence donnée. La variabilité inter-locuteur, quant à elle, regroupe les prononciations différentes d'un même énoncé qui sont le fait de locuteurs distincts.

2. Formalisme de représentation pour le traitement automatique de la parole

Dès lors qu'on aborde la mise en oeuvre d'un système de règles de réécriture formalisant les connaissances que l'on a sur la prononciation des mots d'une langue donnée, tant dans une perspective phonétique que phonologique, plusieurs questions doivent être abordées quant au formalisme adopté et à la nature des règles (contenu, organisation, fréquence d'application, etc.).

2.1. Formalisme

D'une façon un peu trop hâtive et réductrice, on peut avoir tendance à assimiler règles de réécriture et règles génératives. Une règle est dite **générative** quand elle opère sur une forme phonémique abstraite, appelée **forme de base** pour dériver les formes correspondantes effectivement prononcées dans la parole : les **formes de surface.** Nous verrons que le concept de "règles génératives", tel qu'il a pu être proposé par l'école chomskyenne (cf. N. Chomsky et M. Halle 1968) n'est qu'une façon parmi d'autres de traiter les données phonétiques et phonologiques. On verra, en reconnaissance notamment, que d'autres méthodes ont pu être choisies. Le choix concernant la nature des règles utilisées dépend des systèmes de reconnaissance dans lesquels les règles opèrent et des objectifs des concepteurs. Notons en guise de préliminaire que plus la taille du vocabulaire et plus le nombre de locuteurs à reconnaître augmentent, plus les difficultés sont réelles et les choix loin d'être évidents. En synthèse de la parole, les problèmes sont différents, puisqu'il s'agit de faire correspondre à une entrée graphémique donnée une et une seule sortie phonétique. L'utilisation des règles génératives ici est donc discutable. En résumé, il ne semble pas évident que des outils linguistiques proposés dans le cadre de réflexions théoriques sur le langage, ou d'études contrastives de différents systèmes linguistiques, soient parfaitement adaptés au traitement automatique de la parole. En effet, l'approche qui sous-tend ce type d'analyse et qui postule l'existence d'un locuteur-auditeur idéal, appartenant à une communauté linguistique homogène, peut difficilement être retenue par le chercheur s'intéressant aux problèmes concrets de la parole, considérée dans sa réalité et sa contingence même. Néanmoins, le formalisme simple et rigoureux des règles génératives a souvent été utilisé par les chercheurs du domaine.

Une règle de réécriture se présente de la façon suivante :

$$A \longrightarrow B \,/(C)\text{---}(D)$$

où l'élément 'A' se réécrit 'B' dans les contextes gauche et droit facultatifs 'C' et 'D'. Encore une fois, l'élément 'A' peut être autre chose qu'une forme phonémique de base, en synthèse par exemple, il peut s'agir d'une entrée gra-

phémique de longueur variable. De la même façon, l'élément 'B' n'est pas nécessairement un phonème ou une suite de phonèmes, il peut s'agir d'un paramètre acoustique par exemple. On définira donc une règle de réécriture comme un processus qui s'applique sur une forme donnée en entrée de nature et de longueur à définir (phonème, graphème, mot, etc.) pour produire en sortie la ou les formes phonétiques correspondantes. Il s'agit toujours de mettre en évidence les conditions dans lesquelles un phonème, entité linguistique fonctionnelle abstraite, voit certains de ses traits modifiés. Un système de règles est l'ensemble de ces éléments qui le constituent.

2.2. Nature des règles

La cohérence interne d'un système de règles repose sur certains principes déterminés une fois pour toutes : le système doit-il être doté de **règles cycliques** ou **non cycliques** ? les règles sont-elles **ordonnées** totalement, partiellement ou présentées de façon aléatoire ? les règles doivent-elles être appliquées selon leur ordre d'apparition ou de gauche à droite de la chaîne à transcrire ?

Un système de **règles cycliques** répète l'application de règles sur une forme d'entrée autant de fois qu'elles produisent un changement dans les formes de sortie : dans un module de transcription graphème-phonème utilisé pour le français, le mot *médecin* par exemple sera réécrit une première fois [medsɛ̃] et une seconde règle d'assimilation consonantique sera déclenchée pour produire la chaîne [metsɛ̃].

Dans un système de **règles ordonnées,** les formes générées diffèrent en fonction de l'ordre dans lequel les règles sont appliquées. Ainsi, les règles :

(1) eau —> [o]
(2) au —> [o]
(3) e —> ø (élision du 'e' caduc)
(4) a —> [a]
(5) g —> [g]
(6) t —> [t]

permettent de phonétiser correctement le mot *gâteau* ; si la règle (4) par exemple était appliquée avant les règles (1) et (2), on obtiendrait une transcription erronée du mot. Aussi toute règle dont la chaîne est un sous-ensemble d'une chaîne d'une autre règle ne doit-elle pas être classée avant cette dernière.

La hiérarchie d'application des règles dépend de la langue sur laquelle ces règles opèrent. Si, dans une langue comme le hollandais, elles peuvent être appliquées selon l'ordre dans lequel elles sont présentées, en français ce type de stratégie ne peut être retenu : les règles ordonnées doivent être déclenchées de gauche à droite de la chaîne à transcrire.

3. Les connaissances phonétiques et phonologiques en reconnaissance automatique de la parole

3.1. Le décodage acoustico-phonétique

En reconnaissance, puisqu'un paramètre acoustique ne correspond pas de façon univoque à un phonème donné, des connaissances sont nécessaires pour effectuer le passage du signal (paramétré en valeurs acoustiques) à sa description en terme d'unités phonétiques. Cette phase de **décodage acoustico-phonétique** est le point-charnière entre le concret (acoustique) et l'abstrait (linguistique) : en l'absence d'informations articulatoires et perceptives, il faut classer et, autant que faire se peut, identifier les données de parole. Plusieurs techniques peuvent être utilisées pour effectuer ce décodage. Certaines contournent la difficulté de l'analyse phonétique en utilisant des méthodes probabilistes (modélisation markovienne) qui fournissent de très bons résultats (cf. K.F. Lee 1989). D'autres exploitent des méthodes de reconnaissance globale et de quantification vectorielle en utilisant des algorithmes de comparaison dynamique (algorithmes qui calculent le meilleur chemin entre une unité à reconnaître et une unité de référence stockée dans un dictionnaire). L'avantage principal de ces méthodes est leur simplicité, elles ne nécessitent pas de structure sophistiquée pour construire les connaissances linguistiques puisque la représentation des données vient directement de la prononciation des locuteurs. Néanmoins, elles sont difficilement envisageables pour la reconnaissance de parole continue multilocuteur. En effet, le temps d'apprentissage préalable à la reconnaissance (cf. *infra*, chapitre 6), la quantité d'informations à stocker et le temps de calcul pour traiter ces informations deviennent vite inacceptables. D'autres méthodes, enfin, sont fondées sur une modélisation explicite de connaissances phonétiques. Elles feront donc l'objet de notre exposé.

Différents choix doivent être effectués lors du décodage acoustico-phonétique : quelle est la meilleure unité de décision pour effectuer ce décodage ? quels indices acoustiques prendre en compte pour paramétrer le signal de parole ? quelles stratégies utiliser pour réaliser la segmentation et l'identification des unités et quels outils informatiques employer ?

3.1.1. Le choix d'une unité de décision

Reconnaître de la parole continue suppose de choisir une **unité de décision** (ou **unité de reconnaissance**) autre que le mot, afin de prendre en compte les phénomènes de co-articulation entre les mots et de manière à rendre l'unité de reconnaissance moins tributaire du vocabulaire à reconnaître. Aussi, les unités de reconnaissance acoustiques sont-elles souvent plus petites que le mot pour se rapprocher de plus en plus du phonème (cf.

Figure II). L'unité de décision peut être une **syllabe**, une **disyllabe** (transition entre deux centres de syllabes consécutifs), une **demi-syllabe** (portion de parole comprise entre la fin et le début de deux syllabes successives), un **polyphone** (en général un diphone), un **phonème** ou un **phone** (unité infra-phonémique facilement localisable). Le choix de tel ou tel type d'unité de décision repose sur différents critères : d'une part se pose le problème d'encombrement mémoire du stockage des données, d'autre part il s'agit de choisir l'unité qui rend compte au mieux des effets de co-articulation sur un contexte plus ou moins proche. Ces deux aspects sont étroitement liés à la langue dans laquelle a lieu la reconnaissance. En français, les variantes contextuelles de part et d'autre d'un centre de syllabe étant limitées, de nombreux chercheurs choisissent comme unité de décision la portion de parole comprise entre deux parties stables de phonèmes consécutifs : le **diphone** (cf. J. Mariani 1983). Quelle que soit la solution retenue, elle ne peut être optimale : le statut linguistique de l'unité choisie peut être assez bien défini, mais il n'est pas toujours facile de l'identifier (c'est le cas de la syllabe ou du phonème) c'est pourquoi on peut être tenté de choisir une unité de reconnaissance plus facilement localisable, au détriment de la cohérence linguistique.

mot graphémique	*émigrante*
mot phonétique	*emigrãt*
phonèmes	*e m i g r ã t*
diphones	*e em mi ig gr rã ãt t*
syllabes	*e m igrãt*
demi-syllabes	*e em mi ig rã ãt t*
disyllabes	*e emi igrã ãt*

FIGURE II
Les différentes unités de décision permettant de segmenter le mot *émigrante*. (cf. J. Mariani 1989)

3.1.2. Les indices acoustiques

Comme on l'a vu, les sons de la parole ont certaines caractéristiques physiques qui vont servir d'indices pour le décodage acoustico-phonétique (cf. Figure III). Ainsi, la présence d'un silence suivi d'une variation brusque du signal est un indice pour le décodage d'une plosive, la courbe de variation du F0 est un paramètre utilisé pour la détection des sons voisés (cf. Calliope 1989). Pour mettre en évidence ces indices, plusieurs techniques de traite-

ment du signal sont envisageables : **la prédiction linéaire** (détermination, à partir d'un signal, des paramètres du modèle du conduit vocal pouvant produire ce signal) et les **techniques d'analyse spectrale** (recherche sur le signal d'informations fréquentielles et notamment des variations formantiques liées aux fréquences de résonance de la cavité buccale) sont le plus souvent employées.

Consonnes voisées vs. consonnes sourdes

- présence du F0
- rapport entre les énergies haute fréquence et basse fréquence
- rapport entre l'énergie basse fréquence et l'énergie totale
- VOT (Voice Onset Time) : durée de l'établissement du voisement au moment de l'explosion (indice utilisé pour détecter les occlusives voisées)

Occlusives

- zone de silence associée à l'occlusion (occlusives sourdes)
- zone contenant de l'énergie dans les basses fréquences liée à la barre de voisement (occlusives sonores)
- saut d'énergie brusque et forte instabilité spectrale associés à l'explosion

Fricatives

- nombre de passages par zéro du signal
- rapports entre énergie haute et basse fréquence
- durée de la friction
- valeur élevée du centre de gravité spectrale

Nasales

- résonnance marquée entre 200 et 300 hertz
- très faible densité de passages par zéro
- stabilité spectrale
- durée

Liquides

- instabilité et variabilité spectrale (indice qui permet d'opposer les liquides aux nasales)
- forte concentration d'énergie au centre du spectre (utilisée pour la reconnaissance de l'allophone « r »)

FIGURE III
Exemples d'indices utilisés pour la reconnaissance des consonnes.

On peut également souligner l'utilisation des modèles auditifs (modèles du système auditif périphérique) pour la reconnaissance des traits phonétiques (cf. J. Caelen 1979).

Enfin, des études portant sur la recherche d'invariance au niveau acoustique ont montré qu'il était possible d'extraire du signal acoustique des invariants corrélés à certains traits phonologiques. Dans ce cadre, les systèmes de classification des phonèmes en terme de **classes naturelles** et de **traits distinctifs** - un phonème est un faisceau de traits (cf. R. Jakobson *& al.* 1963) - peuvent être utilisés : on recherche l'ensemble des indices acoustiques susceptibles de représenter un trait. Néanmoins, il est nécessaire de souligner le caractère abstrait de tels systèmes : les invariants ainsi dégagés, bien qu'ils se présentent comme des universaux (on les retrouve dans la majorité des langues) ne semblent pas exister dans l'organisation physique du signal de parole. La thèse de l'invariance fait donc l'objet de nombreuses polémiques chez les chercheurs : il semble nécessaire de dégager des descripteurs qui correspondent mieux à la réalité acoustique (cf. J.S. Liénard 1989 ; M. Rossi 1989).

3.1. 3. Segmentation et identification des unités

Segmenter un signal de parole, entreprise apparemment simple, soulève des problèmes de fond :

- Quels indices utiliser pour repérer les frontières de segments et notamment les frontières de mots ? En effet, outre l'imbrication des sons qui composent le signal de parole, les silences rencontrés ne peuvent pas être associés de façon univoque à des séparateurs. Par ailleurs, contrairement aux textes écrits où les blancs indiquent souvent une frontière de mots, en parole continue les mots sont prononcés de façon enchaînée, ce qui pose de difficiles problèmes de segmentation.
- Comment peut-on distinguer les variations porteuses d'informations linguistiques des variations purement physiologiques et sans intérêt ?
- Les phases de segmentation et d'identification doivent-elles être simultanées ou effectuées successivement ?

Ces questions conditionnent la méthode d'analyse ainsi que le modèle de segmentation choisis. En général, la segmentation du signal est préalable à la phase d'interprétation et d'étiquetage ; cette dernière se décompose en deux temps : la détection d'indices acoustiques robustes (identification des macro-classes "consonnes", "voyelles") puis l'identification des différentes classes de voyelles et de consonnes.

Parmi les diverses stratégies utilisées pour l'étiquetage du signal de parole, les **systèmes experts en lecture de spectrogrammes** ont pendant un temps suscité l'enthousiasme des chercheurs. Les connaissances que l'expert phonéticien met en oeuvre pour la lecture de spectrogrammes sont, en effet, extrêmement riches : utilisation simultanée de l'axe temporel et de l'axe com-

binatoire pour effectuer un choix, mémorisation des contraintes phonotactiques de la langue et des variations allophoniques, vision globale et détaillée de la matière phonique, recours à différentes stratégies de décodage (de la gauche vers la droite ou par îlot de confiance). Cet état de fait explique l'engouement général pour l'intelligence artificielle dans les années 80 et plus particulièrement pour le développement de systèmes experts en lecture de spectrogrammes (cf. D. Memmi *& al.* 1983) : on utilise un ensemble de règles de décision contenues dans une base de connaissances (cf. Figure IV). Néanmoins, la mise en oeuvre de tels systèmes, qui passent par la modélisation du raisonnement de l'expert, pose certaines questions classiques : l'expert a-t-il une connaissance exhaustive ? peut-il formaliser toutes ses connaissances ? les règles produites sont-elles toutes pertinentes ? peut-on reproduire fidèlement le raisonnement de l'expert ? Dans l'état actuel des connaissances, aucune réponse satisfaisante n'a encore été trouvée, ce qui explique le déclin de telles stratégies au bénéfice des approches markoviennes, qui ont fait leurs preuves. Par ailleurs, on assiste, depuis 1985 environ, à une approche par **réseaux de neurones** (cf. L. Bottou 1991). On a constaté, en effet, que l'approche neurobiologique semblait bien adaptée à des tâches de bas niveau : les modèles connexionnistes ont besoin de peu de connaissances *a priori* pour classer les phonèmes et peuvent être utilisés comme classificateurs statistiques. En outre, il ne s'agit pas simplement de nouvelles techniques permettant de classer des données, ces modèles offrent également une nouvelle représentation de la connaissance. Enfin, soulignons que l'étiquetage manuel par un expert phonéticien permet de privilégier l'interactivité entre l'expert et les données acoustiques. On dispose ainsi de

- Si le segment à identifier comprend 4 formants, alors il s'agit d'une voyelle.

- S'il existe une zone de concentration du bruit comprise entre 4500 et 4600 hertz et que le phonème [i] est une solution très vraisemblable des segments adjacents, alors il s'agit peut-être du phonème [S] ou [z].

- Si le segment contient du bruit, que sa position n'est pas finale et que la limite basse du bruit est descendante dans la partie droite de ce segment, le segment à droite du segment considéré est probablement une voyelle arrière et peut être un [w].

- Si le second formant est supérieur à 1500 hertz, il ne s'agit probablement pas du phonème [w].

- S'il est possible que le segment soit le phonème [n] et que ce segment soit en fin de groupe rythmique, alors ce n'est vraisemblablement pas un [n].

FIGURE IV
Systèmes experts et décodage acoustico-phonétique : exemple de règles.

documents de référence précis et complets (bases de données de parole segmentées et étiquetées) disponibles pour les études fondamentales sur la parole et utilisables pour une validation des procédures de segmentation et de reconnaissance (cf. R. Carré *& al.* 1984).

3.2. Phonologie et reconnaissance de la parole

Rappelons que dans un système de reconnaissance, la formalisation des contraintes phonologiques consiste à associer à chaque mot graphémique du vocabulaire à reconnaître un ensemble d'informations permettant le passage d'une représentation phonémique abstraite à sa réalisation acoustique. Les informations phonologiques stockées dans un lexique qui contient également des informations morpho-syntaxiques et sémantiques supposent l'élaboration de règles de transcription graphémique-phonémique.

Si, dans une optique orientée synthèse, un module de transcription graphème-phonème est déterministe (il propose une et une seule sortie phonémique pour une entrée graphémique donnée), ce déterminisme n'a pas lieu d'être en reconnaissance. En effet, étant donné qu'il est impossible d'imposer à un utilisateur de technologie vocale une prononciation standard, un système de reconnaissance doit être doté des moyens nécessaires permettant la prise en compte des variantes de prononciation. Le lexique d'un système de reconnaissance doit donc contenir l'ensemble des informations relatives aux altérations phonologiques, permettant le passage d'une représentation graphémique abstraite à sa réalisation acoustico-phonétique. Pour ce faire, plusieurs méthodes peuvent être envisagées : dans le cadre d'une application utilisant un vocabulaire limité, on peut se contenter de lister toutes les prononciations possibles d'un même mot ; en revanche pour la reconnaissance d'un grand vocabulaire, cette stratégie est inexploitable, surtout quand on considère les phénomènes de co-articulation en frontière de mots. Aussi, une autre approche est d'adjoindre au lexique des mots graphémiques, une ou plusieurs formes phonologiques de base à partir desquelles des règles de transduction permettent de dériver les altérations phonologiques les plus fréquentes. On distingue ici deux classes de lexiques :

- dans les **lexiques précompilés**, l'application de **règles génératives** sur les formes de base engendre automatiquement un réseau lexical qui contient les différentes formes possibles d'un mot de façon arborescente ;

- dans les **systèmes à base de règles**, l'application des règles sur un mot s'effectue en cours de traitement par le déclenchement de règles spécifiques. De telles règles, qui analysent l'événement acoustique pour retrouver sa structuration sous-jacente sont appelées **règles analytiques**.

Si les règles génératives permettent de dériver uniquement les occurrences correctes d'un mot, elles augmentent de façon non négligeable la taille du lexique et la phase de précompilation est coûteuse en temps. En revanche,

dans un système de règles analytiques, la précompilation n'a pas lieu puisque les règles sont appliquées durant la reconnaissance, et l'encombrement mémoire nécessaire au stockage des données est réduit. Cependant d'autres problèmes apparaissent : le temps de reconnaissance augmente de façon non négligeable et le risque de produire des prononciations erronées est loin d'être écarté. Le choix des règles (génératives ou analytiques) dépend entièrement des objectifs visés. Dans certains systèmes les deux types de règles sont intégrés. Les règles analytiques, décrivant les procédés acoustico-phonétiques, sont appliquées à des séquences entachées d'erreurs pour la vérification de mots. Elles prennent donc en compte les élisions, les insertions et les substitutions de segments (un même spectre peut être associé à des phonèmes différents). Les règles génératives sont appliquées sur un ensemble de formes de base répertoriées dans un dictionnaire, elles décrivent les mécanismes phonologiques liés à la co-articulation et fournissent un ensemble de formes dérivées. Aux Etats-Unis, les concepteurs du système de reconnaissance HARPY ont mis l'accent sur l'échec de leur méthode initiale, qui avait été de passer par l'intermédiaire de formes phonologiques de base pour effectuer la correspondance graphème-phonème. Dans une seconde étape, ils ont développé les graphes phonologiques directement à partir des entrées lexicales (cf. W.A. Lea 1980). En France, le lexique du système de reconnaissance de la parole continue ESOPE développé au LIMSI à Orsay (cf. J. Mariani 1982) utilise un tel type de représentation : les règles d'un module de transcription graphème-phonème, initialement développé pour la synthèse à partir du texte, ont été modifiées afin de prendre en compte les erreurs de segmentation du système de reconnaissance (cf. F. Néel *& al.* 1986). Les règles du nouveau programme sont appliquées à chaque mot du lexique pour obtenir les variantes de prononciation correspondantes.

Enfin, soulignons le développement, en France, de **bases de données lexicales** en vue d'applications informatiques, dont la reconnaissance de la parole. Ainsi, au CERFIA à Toulouse, G. Pérennou dirige le développement de la base de données grammaticale et phonologique BDLEX (cf. G. Pérennou et M. de Calmès 1986). Concernant plus particulièrement les connaissances phonologiques, les règles sont ordonnées et non cycliques. Le modèle sous-jacent à cette base de données phonologiques est la phonologie générative qui doit permettre de répondre aux objectifs que se sont fixés les chercheurs : traduire les mécanismes phonologiques nécessaires pour prédire la prononciation en notation large. Cette constitution de connaissances communiquées par des experts en phonologie peut ensuite être adaptée aux problèmes de la reconnaissance et de la parole. A chaque entrée lexicale sont associées une forme phonologique de base et des informations morphologiques. En outre, des diacritiques permettent de repérer des variantes individuelles dans la représentation lexicale. Ainsi, le marqueur 'X°' indique qu'un phonème, non prononcé en théorie, peut être actualisé par certains locuteurs, c'est le cas du 'l' final de *persil* ; le caractère X- correspond au caractère fixe

ou mixte d'un phonème selon les habitudes langagières des locuteurs et l'entourage phonétique (le 'q' de *cinq* peut être prononcé différemment dans *cinq maisons* et *il y en a cinq*). Enfin, BDLEX doit offrir un cadre permettant aux phonologues de modéliser un parler régional ou, le cas échéant, un idiolecte spécifique. Ainsi, à partir des règles standard, un autre ensemble de règles sur le parler marseillais est en cours de développement.

Les lexiques électroniques réalisés au LADL dans le même esprit, disposent également d'un dictionnaire phonologique, le dictionnaire DELAP, composé de 500.000 formes (cf. E. Laporte 1988), et présentant l'avantage de contenir des informations précieuses sur les mécanismes phonologiques de la langue française (phénomènes d'assimilation consonantique, élision du 'e' caduc, diérèse ou synérèse du yod, etc.).

4. Les connaissances phonétiques et phonologiques en synthèse automatique de la parole

Nous considèrerons successivement l'utilisation, en synthèse de la parole, des connaissances phonétiques, puis des connaissances phonologiques.

4.1. Phonétique et synthèse de la parole

Etant donné l'aspect continu de la parole et les nombreux effets de co-articulation qu'on y rencontre, se contenter de stocker les ondes naturelles du signal correspondant à chaque phonème isolé d'une langue, et les concaténer pour produire un signal de parole continue est une entreprise vouée à l'échec, puisqu'il faut tenir compte des contraintes de co-articulation entre les sons et entre les mots : les transitions entre les phonèmes, conséquences acoustiques des mouvements de l'appareil vocal, doivent être reproduites pour générer une parole synthétique intelligible. Il est donc nécessaire de disposer d'informations sur les caractéristiques acoustiques des sons de la parole prononcés en continu. Pour ce faire, deux techniques peuvent être utilisées : la synthèse par règles et la synthèse par concaténation d'unités pré-enregistrées.

La synthèse par **règles** consiste à reconstituer le squelette phonétique de la parole en modélisant les transitions entre phonèmes sous forme de règles qui déterminent automatiquement l'évolution des différents paramètres acoustiques indispensables à la commande d'un synthétiseur, tels que l'évolution des formants qui sont définis par un certain nombre de valeurs cibles, les caractéristiques acoustiques des bruits, etc. (cf. Figure V). L'intérêt d'une telle méthode est qu'elle rend possible la commande d'un synthétiseur à partir d'un nombre réduit d'informations (seules les valeurs cibles doivent être

r	—>	avampl = 28 db avtrans1 = 10 ms avtrans2 = 10 ms afampl = 00 db trans1 = 10 ms trans2 = 10 ms f1trans1 = 10 ms f1trans2 = 10 ms/[+vois, +cons]---{r}{r}{r}

Avec

'av' = amplitude de voisement
'af' = amplitude de bruit
'trans1' = transition à gauche des formants F2, F3, F4
'trans2' = transition à droite des formants F2, F3, F4
'f1trans1' = transition à gauche du premier formant
'f2trans2' = transition à droite du premier formant
{r}{r}{r} = 3 phases acoustiques du 'r' à battements décomposable en 4 objets acoustiques
'ms' = milliseconde
'db' = décibel

FIGURE V
Exemple de règles phonétiques utilisées dans un synthétiseur à formants

stockées) ; elle est bien adaptée au synthétiseur à formants ; la description paramétrique imposée permet *a priori* de varier la voix synthétisée ; enfin, cette méthode permet d'améliorer les connaissances des chercheurs sur les mécanismes articulatoires et leurs conséquences acoustiques. Néanmoins, il faut tenir compte des variations acoustiques des phonèmes, déterminées par leur environnement. Pour ce faire, plusieurs centaines de règles, difficiles à élaborer, sont nécessaires ; la majeure partie du travail est à refaire si l'on change de voix et *a fortiori* de langue.

La méthode de synthèse par **concaténation d'unités pré-enregistrées**, développée dans les années 50, permet de réduire le nombre de règles à spécifier pour décrire les évolutions du conduit vocal. Elle consiste à stocker des segments de parole pré-enregistrés qui sont ensuite concaténés pour produire des phrases. Les problèmes à résoudre ici sont de deux sortes : quelles unités choisir et quelle technique mettre en oeuvre pour traiter ces unités, sachant qu'il doit être possible de modifier les paramètres prosodiques sans dégrader le timbre résultant, et comment traiter les discontinuités inévitables aux points de concaténation ? En général, les unités choisies sont des diphones, à savoir les transitions entre deux parties stables de phonèmes. Les mots sont ensuite reconstitués par une concaténation de diphones. L'avantage de cette

méthode est qu'elle ne nécessite pas la mise au point de règles, longue et difficile. Elle peut être utilisée quel que soit le type de synthétiseur développé. Cependant, la qualité de la parole obtenue se fait au prix d'un volume de stockage important (la qualité de la parole synthétique est corrélée au débit de codage choisi). Par ailleurs, l'influence d'un phonème peut s'étendre au-delà de son contexte immédiat ; aussi, pour supprimer les défauts perceptifs résultant des discontinuités aux points de concaténation, et afin d'améliorer la qualité de la synthèse obtenue avec une telle méthode, il peut être nécessaire de disposer de segments encore plus longs que les diphones pour certaines combinaisons de sons difficiles à réaliser. Les systèmes de synthèse en langue française commercialisés aujourd'hui utilisent en général la méthode de synthèse par diphones (cf. F. Emerard 1977).

4.2. Phonologie et synthèse de la parole

Il existe un certain flou terminologique chez les chercheurs en traitement automatique de la parole pour désigner la phase de transcription graphème-phonème indispensable à la synthèse à partir du texte. Pour les uns, il s'agit de traitement linguistique au sens le plus générique du terme, pour les autres de traitement phonétique. Nous parlerons quant à nous de traitement phonologique pour bien le distinguer des autres traitements linguistiques (syntaxique et prosodique) d'une part, à des fins méthodologiques d'autre part : il est essentiel de bien saisir qu'il s'agit d'un processus purement abstrait maniant deux jeux de symboles linguistiques - les symboles de l'écriture et les symboles phonétiques - en spécifiant leur correspondance.

La transcription graphème-phonème - ou transcription orthographique-phonémique - utilisée en synthèse à partir du texte et qui permet de définir des règles pour traiter les phonétisations courantes, mais également les transcriptions ambiguës (par exemple la chaîne *portions* se transcrit /pɔrtjɔ̃ / ou /pɔrsjɔ̃/) n'est pas toujours indispensable : pour la synthèse par concepts, cette étape n'a pas lieu d'être. Ce type de synthèse en effet est interfacé avec un système de génération de texte qui manipule les emplois des mots sous leur forme non fléchie et calcule en même temps la forme fléchie d'un lexème et sa forme phonétique (par exemple le nom *portion* mis au pluriel donne en représentation graphémique *portions* et en représentation phonétique *pɔrsjɔ̃)* (cf. L. Danlos *& al.* 1986). La transcription graphème-phonème met en jeu différentes étapes : prétraitement du texte, transcription proprement dite, vérifications.

Le **prétraitement** d'un texte permet d'effectuer certaines opérations préliminaires à la transcription (traitement des sigles, des abréviations, des nombres, repérage des mots et des signes de ponctuation, etc).

La **transcription** graphème-phonème proprement dite peut se faire par le biais d'un lexique, par l'application de règles ou, le plus souvent, par une combinaison des deux, les exceptions aux règles générales étant listées dans

le lexique. La transcription est généralement guidée par une analyse syntaxique automatique qui permet de résoudre certaines difficultés de transcription associées aux ambiguïtés graphémiques comme les chaînes homographes hétérophones (la chaîne *amer* n'a pas la même représentation phonémique dans *entamer* et *il est amer*).

Enfin, les procédures de **vérification** sont essentielles pour tester la performance du module de transcription graphème-phonème, mettre en évidence les dysfonctionnements et, le cas échéant, améliorer la transcription. En effet, la transcription graphème-phonème d'un texte quelconque reste encore souvent entachée d'erreurs (transcription incorrecte des noms propres, des mots d'emprunt, des sigles et abréviations notamment). Les capacités de prononciation d'un texte par une machine sont limitées par sa capacité de "compréhension" du message.

Cette opération de transcription graphème-phonème, que nous réalisons sans effort quand nous lisons un texte à haute voix, présente des difficultés certaines pour un ordinateur, difficultés qui varient d'une langue à une autre. La difficulté principale tient au fait qu'une langue a plusieurs origines, qu'elle évolue avec le temps et que les règles d'équivalence graphémique-phonémique sont dans certaines langues très difficiles à définir. Si un nombre limité de règles peut suffire pour transcrire un texte en italien, d'autres langues, comme le russe ou l'espagnol, sont également facilement analysables quant à la correspondance graphémique-phonémique, mais posent de réels problèmes eu égard à l'accent lexical qui a une fonction distinctive. Pour d'autres langues comme le français, la transcription est une procédure complexe, qui n'a pas reçu jusqu'à ce jour de solution totalement satisfaisante. Ainsi, en français, il n'y a pas de correspondance simple et directe entre une chaîne de graphèmes et le (ou les) phonème(s) qui lui correspond(ent) : un graphème peut avoir plusieurs représentations phonémiques ('s' –> [s] ou [z]), un phonème peut correspondre à différents graphèmes ([s] –> 't', 's' ou 'c'), dans certains cas plusieurs lettres sont transcrites par un seul phonème ('ai', 'aî', 'ê', 'è', 'e' –> [ɛ]), enfin, il arrive qu'il n'y ait aucune correspondance entre graphèmes et phonèmes ('oiseau' –> [wazo]).

Différentes stratégies peuvent être choisies pour la réalisation d'un programme de transcription graphème-phonème. L'exploitation simultanée d'un ensemble de règles et d'un dictionnaire d'exceptions est couramment utilisée : chaque mot d'entrée en partant de la gauche est comparé aux entrées lexicales d'un dictionnaire d'exceptions. Si à ce stade la transcription n'a pas lieu, en d'autres termes si le mot à transcrire n'est pas stocké dans le dictionnaire, il peut subir certains traitements intermédiaires ou être transcrit directement par l'application de règles générales (cf. Figure VI).

Le découpage morphologique est un exemple de traitement intermédiaire : un mot non codé dans le dictionnaire d'exceptions est découpé en ses différents constituants morphologiques et la racine est de nou-

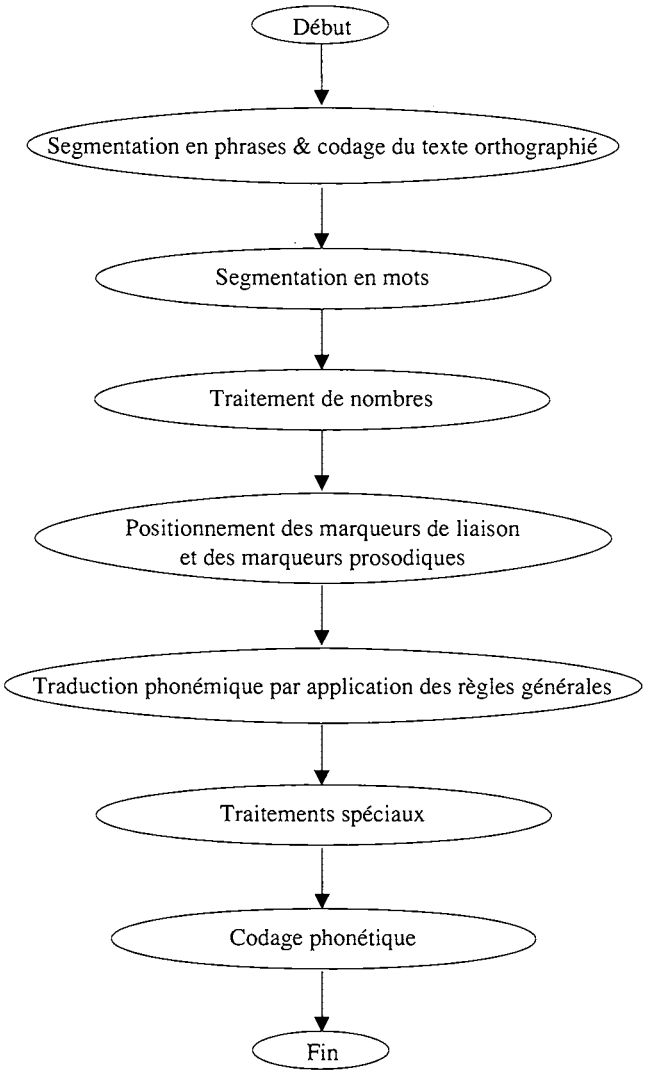


FIGURE VI
Le module de transcription graphème-phonème "Graphon" développé pour la synthèse à partir du texte (cf. B. Prouts 1980)

veau comparée aux mots du dictionnaire. Si la transcription n'est toujours pas effectuée, des règles de réécriture sont déclenchées pour fournir la prononciation correcte du mot. Cette approche par découpage morphophonologique est souvent utilisée pour des langues telles que l'anglais ou l'allemand, où l'organisation morphologique des mots perturbe les règles de prononciation classiques et rend la transcription difficile. En français, l'influence de la morphologie sur la prononciation a pu également être soulignée, par exemple, la règle : "'s' se réécrit [z] dans un contexte vocalique" ne peut s'appliquer aux formes *soubresaut, entresol, parasol,* etc. Cependant, des phénomènes de ce type restent marginaux, une décomposition morphologique préalable à l'application des règles ne semble donc pas nécessaire, il est plus économique de considérer de telles irrégularités comme des exceptions aux règles générales.

Un des premiers programmes de transcription graphème-phonème développé pour le français et associé à un système de synthèse par diphones, reposait sur le principe suivant : chaque mot à transcrire était dans un premier temps comparé à un dictionnaire d'exceptions réduit (20 mots), si le mot ne correspondait à aucun des mots du dictionnaire, il était converti par l'application de règles contextuelles. Ce programme, qui contenait environ 200 règles a été modifié par la suite, les règles ont été augmentées de façon significative (environ 1000 règles) afin d'obtenir un taux de transcription plus correct et pour répondre aux objectifs des chercheurs, qui devenaient plus ambitieux (cf. B. Prouts 1980).

L'approche fondée sur l'utilisation exclusive d'un dictionnaire peut également être envisagée. Ainsi, au CERFIA à Toulouse, le module de transcription graphème-phonème, qui a été développé au début des années 80, reposait sur le principe suivant : la sortie phonémique était obtenue à partir d'une recherche dans un lexique de 2000 mots, chaque entrée du lexique était composée d'un mot graphémique, de son écriture phonémique et de sa catégorie syntaxique (cf. G. Tep 1979). Un tel formalisme rend possible une représentation immédiate de la syntaxe, nécessaire pour résoudre certains problèmes de transcription, tels que la réalisation de la liaison (optionnelle, interdite ou obligatoire). Néanmoins, un analyseur syntaxique peut également être adjoint à un système de règles. Par ailleurs, des problèmes d'encombrement mémoire sont associés au développement d'un lexique, surtout quand il s'agit de traiter un grand vocabulaire. En outre, si le dictionnaire semble être une solution possible pour la conversion des mots les plus courants de la langue, il reste à régler les cas de conjugaison, de formes fléchies, qui peuvent être directement traités par des règles. Enfin, les phénomènes de néologie lexicale ne peuvent être traités qu'au prix d'une mise à jour rigoureuse du dictionnaire.

Une dernière approche peut être fondée sur l'analyse de la structure syllabique : le texte à transcrire est découpé en syllabes, celles-ci sont compa-

rées à des listes de référence pour effectuer la transcription. Ici, soulignons le programme de N. Catach et son équipe (cf. N. Catach 1984), bien qu'il n'ait pas été développé spécifiquement pour la synthèse à partir du texte mais en vue de recherches fondamentales sur le fonctionnement de la langue française. Pour N. Catach, un programme de syllabation automatique rend le texte à traiter beaucoup plus clair, elle y voit au moins quatre avantages : meilleur traitement des voyelles ouvertes et fermées (*porter, dépoter*), meilleur traitement des voyelles nasales, puisqu'une fois le texte syllabé, les voyelles nasales sont reconnues sans l'aide d'aucune règle (*sul-tan, sul-tane*), meilleur traitement des semi-voyelles : les groupes 'cl', 'cr', 'bl' 'br', etc. font toujours partie de la même syllabe, leur présence entraîne donc l'apparition d'un [i] supplémentaire devant le yod (*replier*), enfin, la reconnaissance plus immédiate des digrammes tels que 'ch', 'ph', etc.

5. Perspectives

Le traitement automatique de la parole a conduit à soulever des questions que la linguistique n'abordait pas et à remettre en question certaines dichotomies fondatrices de la linguistique contemporaine (telles que l'opposition saussurienne "langue / parole" ou l'opposition chomskienne "compétence / performance"). Il semble notamment nécessaire de repenser le problème général de la **variabilité** : le postulat selon lequel d'une part la variabilité est reléguée au domaine de la parole, lieu même du contingent, de l'accidentel et de l'hétérogène, d'autre part la variabilité est un bruit perturbateur qui vient dégrader le signal de parole, ne peut être retenu. En effet, la langue n'est pas un matériau inerte, c'est un système de virtualités que les locuteurs actualisent selon les exigences du discours (cf. M. Rossi 1989). Il existe donc une variabilité inhérente au fonctionnement de la langue, qui affecte tous les niveaux du discours (prosodique, phonétique, phonologique, lexical, syntaxique et sémantique). Pour les experts, une étude approfondie des phénomènes de variabilité, de leur mode de fonctionnement et des relations structurelles qu'ils entretiennent entre eux devrait conduire à une meilleure compréhension des mécanismes d'encodage et de décodage de la parole.

Par ailleurs, à côté de la variabilité liée à la langue et à son fonctionnement, il existe dans l'actualisation de la parole, une variabilité inhérente au locuteur qui, loin d'être un bruit perturbateur et aléatoire, donne des indications sur les caractéristiques intrinsèques des sujets parlants. Ces considérations sont de première importance en traitement automatique. En effet, si aujourd'hui en synthèse on arrive à peu près correctement à transcrire un texte de façon standard, cette transcription et les informations prosodiques qui lui sont associées rendent peu compte des aspects variables de la parole. La formalisation des phénomènes de variabilité liés aux différents débits et

styles de parole, ainsi qu'aux idiomes régionaux, devrait permettre d'atteindre cet idéal pour les chercheurs, qui est celui d'une synthèse naturelle, variée et agréable. Ici, les travaux en phonétique et en phonologie jouent un rôle de premier plan : il est nécessaire d'analyser les stratégies phonologiques individuelles des locuteurs, et la cohérence interne de ces stratégies afin de pouvoir les reproduire et pour conduire à une meilleure analyse de la pertinence des règles utilisées, de leur validité et de leur impact dans un système de synthèse. En reconnaissance, des connaissances plus approfondies sur les facteurs extra-linguistiques conditionnant la variabilité rencontrée en parole devraient permettre de mieux cerner le champ de variabilité d'un locuteur, qui n'est pas infini, et de développer des modèles plus appropriés à la reconnaissance de grands vocabulaires multilocuteurs.

Anne LACHERET-DUJOUR
(Université de Caen, LIMSI-CNRS et ELSAP-CNRS)

Repères bibliographiques

[Le domaine de la parole étant tout à la fois extrêmement spécialisé et vaste, relevant de disciplines *a priori* aussi éloignées que la linguistique, la physique, l'électronique, etc. nous avons autant que faire se peut évité de faire appel à des références bibliographiques trop spécialisées ; néanmoins, elles nous ont paru parfois indispensables pour un bon éclairage du texte. Par ailleurs, nous avons essayé de limiter autant que possible les références étrangères.]

PHONÉTIQUE

1. Introduction à la phonétique :

CARTON, F. (1974) : *Introduction à la phonétique du français,* Paris, Bordas.

[Ouvrage de référence en phonétique française, d'accès facile.]

JAKOBSON, R., *& al.* (1963) : *Preliminaries to Speech Analysis : the Distinctive Features and their Correlates,* Cambridge, MIT Press.

[Ouvrage de référence en phonologie où l'on trouvera une présentation des principes de la description phonologique structuraliste ; les classements en terme de traits articulatoires y sont présentés et validés phonétiquement.]

2. Variabilité et invariants phonétiques :

LIENARD, J.S. (1989) : Variabilité, contraintes et spécification de la parole : un cadre théorique, Actes du Séminaire : *Variabilité et spécificité du locuteur : études et applications*, Marseille, Société Française d'Acoustique, 1-10.

[Article de fond qui présente de nouveaux paradigmes de recherches en traitement automatique de la parole, concernant les aspects non linguistiques du signal de parole.]

ROSSI, M. (1989) : De la quidité des variables, Actes du Séminaire *Variabilité et spécificité du locuteur : études et applications,* Marseille, Société Française d'Acoustique, 11-31.

[Présentation des grands types de variabilité en parole, et proposition d'une méthodologie pour leur codage.]

3. Décodage acoustico-phonétique :

BOTTOU, L. (1991) : *Une approche théorique de l'apprentissage connexionniste ; applications à la reconnaissance de la parole.* Thèse de l'Université de Paris Sud.

[Thèse présentant une application de décodage acoustico-phonétique en reconnaissance de la parole au moyen de réseaux neuronaux.]

CAELEN, J. (1979) : *Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique,* Thèse de Doctorat d'Etat, Toulouse.

[Thèse d'accès difficile présentant le développement d'un modèle auditif pour la reconnaissance de traits phonétiques.]

CALLIOPE (1989) : *La parole et son traitement automatique,* Paris, Masson.

[Ouvrage d'introduction à la communication parlée qui fournit un bon panorama de l'état actuel du domaine. Rassemblant une bonne partie des recherches françaises en traitement automatique de la parole, il constitue une référence indispensable pour les chercheurs tant débutants que confirmés.]

CARRÉ, R., *& al.* (1984) : The French Language Database. Defining, Planning and Recording a Large Database, *IEEE International Conference on Acoustic Speech and Signal Processing*, vol 3, San Diego, The Institute of Electrical and Electronics Engineers Signal Processing Society, 42.10.1- 42.10.4.

[Article présentant une base de données des sons du français, développée sous l'égide du groupe Communication Parlée de la Société Française d'Acoustique et utilisée en France dans de nombreux domaines du traitement automatique de la parole.]

LEE, K.F. (1989) : Hidden-Markov Models : Past, Present and Future, dans J.P Tubach & J. Mariani (eds.) : *European Conference on Speech Communication Technology*, vol 1, Paris, 148-155.

[Article de fond présentant l'état de l'art en décodage acoustico-phonétique par l'application de modèles markoviens.]

MARIANI, J. (1983) : Reconnaissance phonétique par diphonèmes, *Speech Communication,* 2, North Holland, The European Association for Signal Processing, 203-206.

[Présentation d'une stratégie de segmentation du signal de parole pour le décodage acoustico-phonétique en reconnaissance automatique.]

MARIANI, J. (1989) : Recent Advances in Speech Processing, *IEEE International Conference on Acoustic Speech and Signal Processing*, Glasgow, The Institute of Electrical and Electronics Engineers Signal Processing Society, 429-440.

[Synthèse de l'état de l'art en traitement automatique de la parole, bibliographie abondante.]

MEMMI, D., *& al.* (1983) : Un système expert pour la lecture de sonagrammes, *Speech communication*, 2, North Holland, The European Association for Signal Processing, 234-236.

[Présentation de l'utilisation de l'intelligence artificielle pour l'étiquetage automatique en reconnaissance de la parole.]

PHONOLOGIE

1. Théories phonologiques :

CHOMSKY, N., & HALLE, M. (1968) : *The Sound Pattern of English,* New York, Harper and Row.

[Ouvrage de référence en phonologie, présentant les principes de la linguistique américaine transformationnelle. L'utilisation de la syntaxe et des règles de transformation comme principe d'explicitation des mécanismes phonologiques y est présentée en détail.]

LÉON, P., *& al.* (1977) : *La phonologie, les écoles et les théories,* Paris, Klincksieck.

[Présentation historique des théories et des concepts fondateurs de la phonologie européenne et américaine au XXe siècle.]

2. Bases de données phonologiques :

LAPORTE, E. (1988) : *Méthodes algorithmiques et lexicales de phonétisation de textes, applications au français,* Thèse de 3ème cycle, Université de Paris 7.

[Présentation d'une stratégie de développement de lexiques phonétiques électroniques en français.]

PÉRENNOU, G., & DE CALMES, M. (1986) : *BDLEX 1,* base de données et de connaissances du français parlé, Actes du séminaire *Lexique et traitement automatique des langages*, Toulouse, Société Française d'Acoustique, 243-258.

[Présentation de la base de données lexicale et phonologique développée sous l'égide du Groupe Communication Parlée de la Société Française d'Acoustique.]

3. Phonétique / phonologie et synthèse de la parole :

DANLOS, L., EMERARD, F., LAPORTE, E. (1986) : *Synthesis of spoken messages from semantic representations (Semantic representation-to-text-to-speech-system)*, Proceedings of International Conference on Computational Linguistics, Bonn, 599-604.

[Article présentant une application de génération automatique de texte à la synthèse de la parole en français.]

EMERARD, F. (1977) : *Synthèse par diphones et traitement de la prosodie,* Thèse de 3ème cycle, Université des Langues & Lettres de Grenoble.

[Thèse présentant la réalisation d'un système de synthèse dans une perspective linguistique (phonétique et phonologique). Un historique des travaux en synthèse de la parole y est également effectué.]

CATACH, N. (1984) : *La phonétisation automatique du français, les ambiguïtés de la langue écrite,* Paris, Editions du CNRS.

[Introduction théorique aux problèmes de la transcription graphémique-phonémique en français.]

PROUTS, B. (1980) : *Contribution à la synthèse de la parole à partir du texte, transcription graphème-phonème en temps réel sur microprocesseur,* Thèse de Docteur Ingénieur, Université de Paris Sud.

[Thèse présentant le développement d'un module de transcription graphème-phonème utilisé pour la synthèse à partir du texte en français.]

TEP, G. (1979) : Système de génération de phrases phonétiques, *10èmes Journées d'étude sur la parole*, Grenoble, Société Française d'Acoustique, 273-283.

[Présentation d'une stratégie lexicale de transcription phonémique utilisée pour la synthèse du français.]

4. Phonologie et reconnaissance de la parole :

LEA, W.A. (1980) : *Trends in Speech Recognition,* New-Jersey, Englewood Cliffs, Prentice-Hall.

[Ouvrage de référence en reconnaissance automatique de la parole ; de nombreux systèmes et stratégies de reconnaissance y sont décrits et commentés par divers spécialistes américains du domaine.]

MARIANI, J. (1982) *ESOPE : un système de compréhension de la parole continue,* Thèse d'Etat, Université de Paris VI.

[Thèse présentant le développement d'un des premiers systèmes de reconnaissance automatique de la parole continue pour le français.]

NÉEL, F. *& al.* (1986) : Module de traduction phonétique avec variantes, Actes du séminaire *Lexique et traitement automatique des langages*, Toulouse, Société Française d'Acoustique, 129-138.

[Présentation de règles phonologiques développées pour le français et utilisées en reconnaissance automatique, en vue de prendre en compte les variantes de prononciation les plus courantes.]

2
PROSODIE

Après un rappel théorique des faits prosodiques envisagés dans une perspective linguistique, nous présenterons les fonctions qu'ils assument en traitement automatique de la parole. Nous verrons comment la prosodie peut être utilisée dans un système de reconnaissance pour segmenter un signal de parole continue (point de départ de l'analyse) en étiquettes acoustico-phonétiques, voire linguistiques (étape finale de la reconnaissance). Quant à la synthèse à partir du texte, on présentera les méthodes utilisées pour segmenter les mots entrés dans un système de synthèse en groupes prosodiques auxquels on associe des paramètres acoustiques adéquats spécifiant la durée, l'accentuation et les variations mélodiques des segments, en vue de produire en sortie un signal de parole intelligible et acceptable pour les auditeurs.

1. Rappels linguistiques

La production d'un message oral implique l'intervention complexe des muscles de la **phonation**, ainsi que le contrôle des différents **paramètres prosodiques** qui constituent l'intonation : variations de hauteur, d'intensité et de durée.

Les variations de **hauteur** dépendent de la fréquence de vibration des cordes vocales : le fondamental mesuré en hertz et noté F0 (cf. *supra*, chapitre 1). Les variations d'**intensité** de la voix lors des vibrations des cordes vocales, sont fonction de l'énergie contenue dans le signal dans un intervalle de temps donné ; l'intensité est étroitement liée au mode d'articulation et au

contexte consonantique : le non voisement et la fermeture des phonèmes augmentent l'intensité des voyelles voisines. Les variations de **durée** affectent les segments phonétiques ainsi que les énoncés prononcés, et incluent les temps de pause.

Les paramètres suprasegmentaux sont liés à l'organisation temporelle de l'énoncé ; le **débit** de parole est caractérisé par deux variables : la vitesse de parole et la vitesse d'articulation. La **vitesse de parole** traduit la **vitesse d'élocution** d'un sujet ; elle s'obtient en divisant le nombre de phonèmes émis par un locuteur par son **temps de locution** (temps passé à prononcer un énoncé). La **vitesse d'articulation** - ou **vitesse de phonation** - s'obtient en divisant le nombre de phonèmes prononcés par le **temps d'articulation** (temps de locution moins temps de pause). Les paramètres suprasegmentaux sont également liés au **rythme** marqué par les phénomènes d'accentuation, et dans une moindre mesure au **timbre**, lié à l'enveloppe spectrale du signal de parole.

Dans la perception de la parole, la fonction principale de la prosodie est d'organiser le signal acoustique en un flux auditif cohérent, structuré rythmiquement et dont l'intonation naturelle contribue à l'identification et à la compréhension de segments phonétiques même dans un milieu bruité. Différentes expériences ont montré qu'une déformation artificielle de l'organisation prosodique d'un énoncé, telle que l'introduction de pauses à des endroits incongrus, ou l'application de variations erronées du F0, perturbait sensiblement la compréhension de cet énoncé.

Malgré le terme "suprasegmental" (terminologie américaine) qui pourrait laisser penser que la prosodie échapperait à toute description linguistique, de multiples études acoustico-phonétiques ont montré que non seulement le continuum prosodique peut être segmenté en formes élémentaires (contours mélodiques types, groupes accentuels, etc.), mais qu'en outre l'observation de ces différents segments prosodiques (ou **prosodèmes**) doit s'inscrire dans le cadre d'une analyse fonctionnelle : comprendre et décrire comment ces formes élémentaires se regroupent pour former des **mots prosodiques** et structurer linguistiquement la parole. Enfin, si la prosodie de la parole est déterminée par de nombreux facteurs relevant du **code linguistique** (phonologique, syntaxique, sémantique), elle est également conditionnée par des phénomènes **extra** et **para-linguistiques** (contexte d'élocution, émotivité du sujet, constitution physiologique des organes de production liée essentiellement à l'âge et au sexe du locuteur). Ainsi, en se fondant sur les trois classes de fonction du langage distinguées par le psychologue allemand Karl Bühler (cf. O. Ducrot et T. Todorov 1972), dans l'acte de parole, l'auditeur peut reconnaître qui parle, sur quel ton il parle et ce qu'il dit. Trois plans de la prosodie sont donc susceptibles d'être définis : le plan **expressif**, le plan **appellatif** et le plan **représentatif**. Ces plans résultent de la volonté du locuteur ou peuvent être utilisés dans la perception de l'auditeur qui s'en sert comme indice de décodage. Le plan expressif concerne tout ce qui permet de repérer l'identité du sujet (ori-

gine géographique, socio-culturelle, âge, sexe). Au plan appellatif, on trouve tous les traits reflétant les caractéristiques psycho-physiologiques d'un locuteur (attitude, émotion, expressivité). Le plan représentatif correspond à tout ce qui, dans la prosodie, participe à la transmission d'informations, c'est-à-dire tout ce qui autorise la reconnaissance de phrases douées de sens.

On comprendra donc aisément que les paramètres prosodiques jouent un rôle en synthèse de la parole mais également en reconnaissance. Les recherches en traitement automatique de la parole nécessitent donc *a priori* des connaissances approfondies sur les mécanismes physiologiques et les contraintes de production influant sur les variations micromélodiques. De la même façon, les contraintes linguistiques à l'origine des variations macroprosodiques doivent être explicitées et modélisées. Ces modélisations peuvent apparaître comme marginales dans le cadre des recherches en reconnaissance de la parole, nous en voulons pour preuve le peu de travaux portant sur le sujet. En revanche, elles sont fondamentales en synthèse. C'est essentiellement sur ce dernier point que nous centrerons notre exposé en montrant comment les recherches en synthèse de la parole ont conduit au développement de modèles et de règles formalisant le lien entre organisation prosodique et structuration linguistique.

2. Conceptions de la prosodie en traitement automatique

Les recherches menées sur la prosodie dans le cadre du traitement automatique de la parole ont été marquées par deux périodes antagonistes. D'abord, pendant les années 70-80, les chercheurs ont essayé de mettre en lumière le lien étroit qui existerait entre la structure prosodique des énoncés et leur structure linguistique, plus spécifiquement syntaxique : "la prosodie participe à la structure syntaxique de la phrase au même titre que l'ordre des morphèmes et que leur désinence" (cf. Ph. Martin 1973). Puis le semi-échec de ces travaux a conduit les chercheurs, dès les années 80, à remettre en question de façon parfois radicale cette congruence entre syntaxe et prosodie et à se tourner vers d'autres voies d'investigation, notamment rythmiques. Sans prendre parti dans le débat mené aujourd'hui sur ce thème, nous montrerons qu'il est sans doute difficile de parler d'isomorphisme entre structures prosodiques et structures syntaxiques, mais que cependant les travaux menés sur la prosodie ne peuvent se faire totalement indépendamment de la syntaxe.

2.1. Isomorphisme entre prosodie et syntaxe

Parmi les pionniers de la recherche sur la prosodie française, nous citerons P. Delattre, qui est l'un des premiers à avoir établi un lien direct entre la

structuration syntactico-sémantique et les variations prosodiques en français. Il a notamment mis en lumière la fonction significative de l'intonation : elle permet de véhiculer les différentes modalités d'une séquence (déclarative, interrogative, parenthétique, exclamative). Il a également montré comment l'organisation accentuelle d'une phrase reflète sa structure sémantique et syntaxique (cf. P. Delattre 1966, 1967).

Dans cette lignée théorique, les chercheurs en traitement automatique de la parole n'ont de cesse pendant la décennie 70-80 de mettre en évidence la corrélation étroite entre syntaxe, variations mélodiques et structuration accentuelle ; pour les uns, le rôle des variations intonatives et accentuelles serait de donner au récepteur d'un message une information sur la structure syntaxique globale de l'énoncé, pour les autres, les variations suprasegmentales ont également une fonction sémantique. Les modèles de congruence se multiplient, proposant des règles qui lient structure prosodique et structure linguistique. En partant de l'analyse syntaxique d'une phrase, on développe des systèmes de marques relationnelles qui segmentent le texte en groupes syntactico-prosodiques, certains même précisent la relation sémantique existant entre ces groupes. Ainsi, au MIT, J. Vaissière tente de définir des patrons intonatifs types dans le contour mélodique d'une phrase à partir d'une représentation arborescente syntaxique simplifiée ; elle développe une grammaire de la prosodie rendant compte des courbes de variation du fondamental et de la durée (cf. J. Vaissière 1975). A l'Institut de Phonétique d'Aix en Provence, M. Rossi définit des indices acoustiques et perceptuels caractéristiques de l'intonation prédicative de la phrase française (cf. M. Rossi *& al.* 1981). Au même endroit, A. Di Cristo développe un ensemble de règles intonosyntaxiques permettant de générer automatiquement les structures intonatives types du français à partir d'une grammaire en constituants immédiats (cf. A. Di Cristo 1975). Toujours à Aix, Ph. Martin propose un ensemble de règles d'accentuation et d'attribution de contours mélodiques fondé sur une grammaire de dépendance (cf. Figure I.A). En outre il définit une matrice de traits prosodiques caractérisant la pente, l'amplitude et la durée des contours pour rendre compte des variations mélodiques ; outre ces règles intonosyntaxiques, une règle d'attribution d'un contour supplémentaire marque la division sémantique de l'énoncé en terme de **thème** et de **rhème** (cf. Ph. Martin 1976). Au CNET, à Lannion, un ensemble de règles intonosyntaxiques développées à partir de connaissances syntaxiques et phonotactiques (longueur des mots) est proposé par D. Larreur & F. Emerard (1977).

Une autre méthode peut également être envisagée : exploiter la ponctuation et une syntaxe rudimentaire afin de positionner des marqueurs prosodiques indépendamment de la fonction syntaxique des syntagmes et de la structure linguistique générale. Ainsi, au LIMSI à Orsay, C. Choppy propose d'utiliser des critères lexicaux (distinction entre mots outils et mots lexicaux) et phonotactiques, ainsi que la ponctuation comme indices des variations

La grande (C3 fm) sœur (C2 FM) de Marie (C1 FMT) écrit (C3 FM) des lettres (CS fm) parfumées (C0)

La grande sœur (C2) de Marie (C1) écrit (C3) des lettres parfumées (C0)

La grande sœur de Marie (C1) écrit des lettres parfumées (C0)

Avec
C1 = court, ample et montant
C2 = court, ample et descendant
C3 = court, restreint et montant
CS = court et neutre
C0 = long, descendant (contour final d'un énoncé déclaratif)
fm = frontière syntaxique mineure
FM = frontière syntaxique majeure
FMT = frontière syntaxique majeure terminale

FIGURE I.A
Génération de contours mélodiques (Cn avec n variant de 1 à 8) à partir de la stucture syntaxique de surface avec désaccentuation partielle. Un contour est un faisceau de traits : ± extrême, ± ample, ± montant (cf. Ph. Martin 1976)

intonatives, indépendamment de la structure syntaxique des groupes ainsi constitués (cf. C. Choppy et J.S. Liénard 1977).

2.2. Le principe d'eurythmie

Le début des années 80 correspond à une remise en question théorique, toujours d'actualité aujourd'hui, fondée en partie sur les limites des applications en traitement automatique de la parole, des modèles de congruence syntaxe / prosodie développés jusqu'alors. Le lien intono-syntaxique fait l'objet de nombreux débats chez les chercheurs et de prises de position antagonistes. Certains pensent que cette corrélation est aléatoire et que la macromélodie relève d'un domaine plus large incluant la sémantique (cf. G. Caelen-Haumont 1991). En effet, à une même structure syntaxique peuvent être associées plusieurs stratégies prosodiques, on ne peut donc pas se fier à une correspondance univoque prosodie / syntaxe pour reconnaître la parole. En synthèse, il faut faire des choix parmi ces variations possibles, choix qui peuvent être guidés par des contraintes non syntaxiques. D'autres proposent de tempérer cette relation. Ainsi, pour D. Hirst "la principale critique que l'on peut formuler à l'encontre des travaux sur l'intonation, c'est cette tentation d'établir un lien trop direct entre acoustique et linguistique" (cf. D. Hirst

1986). D'autres enfin, ne remettent pas en cause les rapports syntaxe / prosodie, mais estiment que les modélisations syntaxiques utilisées habituellement, qui reposent le plus souvent sur des grammaires en constituants immédiats, sont inopérantes pour rendre compte de ces rapports et des mécanismes syntaxiques à l'oeuvre dans la communication parlée. Il est donc nécessaire de proposer de nouveaux modèles syntaxiques en adéquation avec la réalité de l'oral.

Par ailleurs, les travaux de P. Fraisse sur la perception et le rythme du temps (cf. P. Fraisse 1967) soulèvent l'intérêt de nombreux chercheurs en parole : le rythme joue un rôle fondamental dans la performance prosodique, aussi bien du point de vue de la perception que sur le plan de la production. La parole en effet n'est pas seulement contrôlée par des contraintes linguistiques, il s'agit d'un phénomène biologique, variable selon la situation de communication et où l'organisation rythmique est essentielle (cf. Ph. Martin 1987).

Enfin, les travaux en **phonologie non linéaire**, et notamment en phonologie **métrique** contribuent à ébranler la naïve certitude de la congruence entre structuration syntaxique et variations prosodiques. Si cette congruence peut être respectée dans des tâches de lecture, elle est aléatoire en situation de dialogue. Quant au lien entre la structure prosodique et l'organisation sémantique d'un énoncé, il est encore plus incertain. Il est nécessaire de prendre en compte d'autres critères pour décrire les phénomènes prosodiques, dans la mesure où les critères linguistiques habituellement utilisés s'avèrent insuffisants.

Le **rythme** est l'un des principaux supports de la description métrique, il est appréhendé ici dans une perspective textuelle : un texte dispose d'une potentialité rythmique liée à sa structure morphophonémique profonde, qui est actualisée dans le cadre de la performance. C'est ainsi qu'est développé le **principe d'eurythmie** qui désigne le processus selon lequel, quand nous parlons, nous avons tendance à découper les groupes prosodiques que nous formons en fonction de contraintes rythmiques (groupes prosodiques équilibrés du point de vue du nombre de syllabes accentuées et inaccentuées qu'ils contiennent). Une nouvelle méthode de structuration prosodique de l'énoncé consiste alors à le segmenter en groupes accentuels à partir de connaissances syntaxiques : "chaque mot porte un accent dont la force est proportionnelle à l'importance de la coupe syntaxique qui suit ce mot" (cf. F. Dell *& al.* 1984). Ces patrons accentuels — ou **grilles métriques** — sont ensuite modifiés selon le principe d'eurythmie qui s'appuie sur des règles de proéminence (alternance d'accents forts et faibles dans une séquence accentuelle). En outre, un groupe accentuel a un certain degré d'acceptabilité lié au nombre de syllabes qui le constituent. Ce principe, fondé sur la compétence sous-jacente d'un locuteur-auditeur, repose sur l'idée qu'une structure accentuelle bien formée correspond à des schémas prédéfinis. Il est adapté pour le français par F. Dell (cf. F. Dell *& al.* 1984) qui propose de répondre à la question suivante : com-

ment définir une grammaire du français standard associant chaîne syllabique et profil mélodique ?

Dans le cadre des recherches en traitement automatique de la parole, Ph. Martin redéfinit ce principe d'eurythmie, en se fondant plus particulièrement sur l'équilibre syllabique des constituants prosodiques : les mots prosodiques dérivés de la structure syntaxique d'une phrase sont réorganisés de telle manière qu'ils forment des séquences acceptables en vertu de leur caractère eurythmique (tendance à la réalisation d'unités présentant le même nombre de syllabes) (cf. Ph. Martin 1987).

Dans l'absolu, cette nouvelle approche permet de s'affranchir de l'analyse syntaxique d'un énoncé, un index de disrythmie pouvant se calculer à partir du nombre de syllabes comprises dans une groupe accentuel. La structure rythmique se définit alors comme une organisation hiérarchique indépendante composée d'unités rythmiques minimales. Dans les faits, cette autonomie n'est que partielle, puisque l'augmentation de durée de certaines structures rythmiques est proportionnelle à l'importance de la frontière syntaxique correspondante. Par ailleurs, une analyse syntaxique est en général l'étape préliminaire à partir de laquelle la structuration rythmique est définie. Ainsi, dans le système de synthèse du français développé à Grenoble (cf. G. Bailly 1986), le texte à synthétiser est fragmenté en mots prosodiques par une grammaire de dépendance qui exprime les relations syntaxiques entre les mots, le message est ensuite fractionné en groupes de phonation séparés par des pauses dont l'occurrence et la durée sont conditionnées par des critères syntaxiques. Ce n'est qu'en dernier lieu que sont déterminés les groupes accentuels en fonction de contraintes syllabiques (cf. Figure I.B.).

Le joli (IT) petit (DG) Chat (DD) noir (ID) boit (DD) du lait

Le joli (IT) petit chat noir (ID) boit du lait

Avec

IT = interdépendance (frontière de mots prosodiques : 4 syllabes au plus)
DH = dépendance gauche
DD = dépendance droite
ID = indépendance (frontière de groupes de phonations : 8 syllabes au plus)

FIGURE I.B
Segmentation de l'énoncé par une grammaire de dépendance exprimant les relations syntaxiques entre les mots lexicaux avec effacement partiel des marqueurs (cf. G. Bailly 1986)

3. Reconnaissance automatique de la prosodie

En reconnaissance de la parole, l'utilisation de la prosodie n'est pas un passage obligé, mais peut être un outil efficace pour limiter l'indéterminisme du processus de reconnaissance automatique d'énoncés soumis à un système. Le découpage suprasegmental d'un énoncé et l'identification de son schéma prosodique devraient permettre de disposer de points d'ancrage fiables pour réduire le nombre de chemins vraisemblables, aussi bien en ce qui concerne l'attribution d'étiquettes acoustico-phonétiques que l'accès au lexique ou l'identification des structures syntaxiques. Les paramètres prosodiques peuvent notamment être utilisés comme indices pour l'identification des unités segmentales. Ainsi, des phénomènes micromélodiques tels que l'absence, la chute ou la continuité du fondamental peuvent servir pour l'identification des consonnes dans un système (pour une explication détaillée cf. Calliope 1989, pp. 431-432). Les variations prosodiques peuvent également servir d'indices pour signaler la structure syntaxique sous-jacente d'un énoncé : l'utilisation de la **ligne de déclinaison** (dans beaucoup de langues, tendance de la ligne mélodique à diminuer de hauteur du début à la fin d'un groupe et d'une phrase) peut aider à repérer les limites des groupes syntaxiques et les fins de phrases. De même, des variations significatives du F0 et du rythme peuvent signaler des frontières lexicales ou syntaxiques. Enfin, la prosodie peut être utilisée comme indice signalant les parties sémantiquement importantes d'un énoncé. Cependant, le développement de règles prosodiques fiables reste une entreprise ardue et les réalisations dans ce domaine sont restreintes. Cette limitation s'explique en grande partie par l'extrême variabilité prosodique inter- et intra-locuteur qui rend difficile l'identification d'invariants prosodiques pouvant être corrélés de façon certaine à des unités linguistiques. La difficulté est d'autant plus réelle que la performance prosodique est également guidée par des contraintes ectolinguistiques difficilement formalisables.

4. Synthèse automatique de la prosodie

En synthèse, la génération automatique de la prosodie consiste à attribuer à un énoncé à synthétiser un contour prosodique adéquat, contribuant ainsi à rendre le signal intelligible et naturel. Ici, notons qu'il est difficile d'établir des échelles de naturel, étant donné le caractère éminemment subjectif du concept. En outre, naturel et intelligibilité n'évoluent pas nécessairement dans le même sens. Ainsi, des expériences ont montré que changer le naturel d'un système de synthèse pouvait avoir des effets pervers sur l'intelligibilité et donc sur la compréhension du message. En revanche, le critère d'intelligibilité est calculable, l'importance de la prosodie pour la segmentation et la

compréhension d'un discours a été démontrée tant au niveau modal qu'au niveau phonétique et lexical.

4.1. Génération des paramètres prosodiques en synthèse

En synthèse de la parole à partir du texte, la génération des paramètres prosodiques nécessite, outre la segmentation de l'énoncé en groupes prosodiques et la génération automatique des pauses, l'attribution d'un profil mélodique et d'une durée spécifique pour chacun des segments ainsi constitués. On verra dans un premier temps, qu'il est nécesaire de disposer d'un module d'analyse morphologique et syntaxique pour effectuer cette opération. Cela n'est pas le cas dans un système de synthèse par concepts, interfacé avec un générateur, puisque comme la phonétique (cf. *supra*, chapitre 1) la syntaxe est calculée automatiquement par le système de génération. Il suffit d'intégrer dans le générateur un module prosodique qui produit des phrases étiquetées de marqueurs syntactico-prosodiques (cf. L. Danlos *& al.* 1985). Nous présentons ensuite une méthode possible de discrétisation du continuum prosodique, étape indispensable à l'application de règles mélodiques, accentuelles et de règles de durée. Les différents modèles utilisables pour la génération automatique de la prosodie sont également discutés. Enfin, nous présentons le rôle joué par l'intensité dans la génération automatique de la prosodie en synthèse à partir du texte.

4.1.1. Les informations syntaxiques

Deux grandes stratégies (à savoir avec ou sans analyse syntaxique) peuvent être choisies pour la génération automatique des paramètres prosodiques en synthèse.

Dans la stratégie **avec** analyse syntaxique, l'analyse morphologique du texte à synthétiser est la première phase nécessaire à la génération automatique de la prosodie. Elle permet de déterminer les mots de la langue et les paramètres linguistiques (variables grammaticales) auxquels ils correspondent. Cette analyse morphologique peut être réalisée de deux façons. La première solution consiste à stocker les mots usuels dans un gigantesque dictionnaire de formes complètes où sont codés les attributs morphosyntaxiques de chaque entrée. Les systèmes fondés sur la consultation de dictionnaires doivent être mis à jour régulièrement afin de traiter les phénomènes de néologie lexicale. La seconde stratégie est beaucoup plus économique, puisque seuls les morphèmes orthographiés sont listés dans un lexique (préfixes, bases, suffixes et désinences). A ces formes orthographiées sont associées des transcriptions phonétiques et une grammaire d'états finis qui décrit, par l'application de règles de transduction, les possibilités d'assemblage de ces morphèmes pour constituer des mots.

La seconde phase du marquage prosodique consiste à lier structure prosodique et structure syntaxique profonde. A ce stade les systèmes de marquage sont confrontés à diverses difficultés liées aux limites des analyseurs syn-

taxiques : ces analyseurs ont une connaissance morphologique partielle de l'énoncé, de nombreuses ambiguïtés morphosyntaxiques conduisent à une génération syntaxique non déterministe (trop grand nombre d'arbres syntaxiques). Les modèles de congruence doivent donc tenir compte de cette contrainte.

Dans la stratégie **sans** analyse syntaxique, le simple balayage d'un lexique où sont associés graphèmes et marques prosodiques permet d'affecter des marqueurs spécifiques aux mots détectés dans la chaîne orthographique à synthétiser.

4.1.2. Les marqueurs prosodiques

Le marquage prosodique correspond à une structure prosodique sous-jacente que l'on peut structurer en six niveaux (cf. Figures II et III). Ces niveaux s'organisent de la façon suivante, indépendamment des règles de désaccentuation partielle ou totale :

- la **phrase**, dont la fonction est essentiellement modale (interrogation, affirmation, exclamation, négation),
- le **groupe de phonation**, qui est lié aux contraintes physiologiques

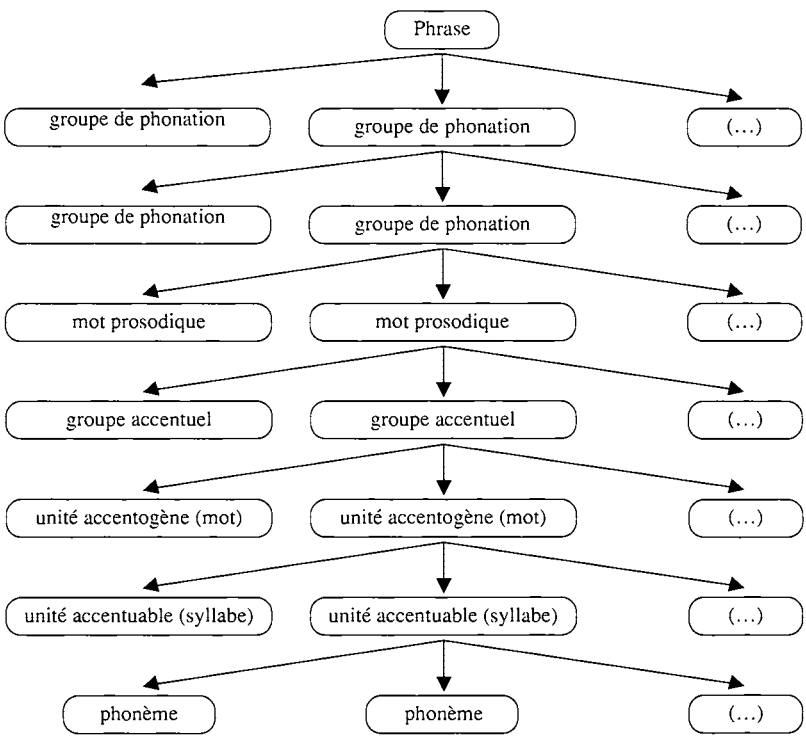


FIGURE II
Les niveaux de structuration prosodique

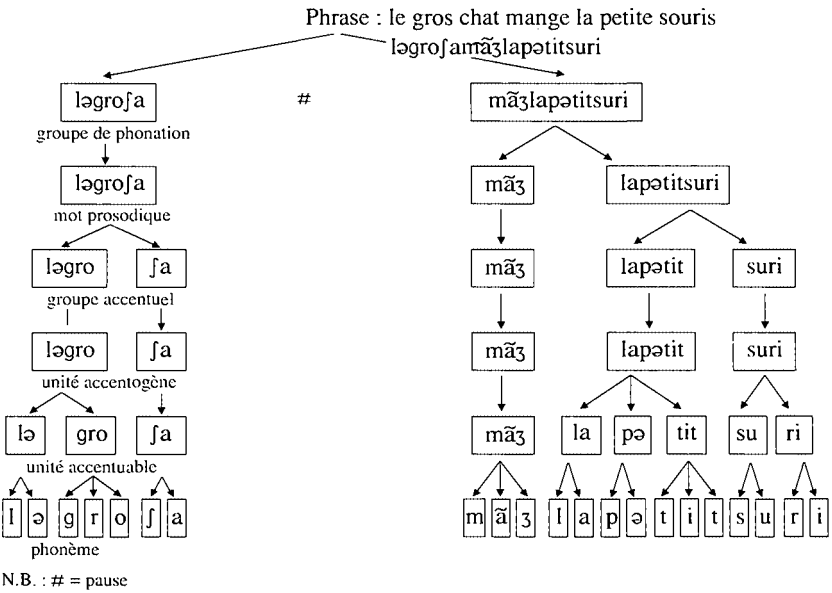


FIGURE III

Segmentation prosodique de la phrase : Le gros chat mange la petite souris.

associées aux poumons — les groupes de phonation sont séparés par des pauses respiratoires,

- le **mot prosodique,** qui correspond à l'intervalle temporel séparant deux syllabes fortement accentuées,

- le **groupe accentuel,** facultatif, qui contient une seule syllabe accentuée — dans certains cas il correspond à l'unité accentogène,

- **l'unité accentogène**, qui est un mot susceptible de recevoir un accent,

- **l'unité accentuable**, qui correspond à la syllabe, unité rythmique par excellence,

- **le phonème**, qui est l'unité microprosodique minimale.

Il est également nécessaire de tenir compte des règles d'accentuation et de désaccentuation partielle ou totale, qui viennent altérer la structure prosodique originelle pour répondre à différents principes syntaxiques et eurythmiques. Ainsi, un mot outil ne sera jamais accentué, deux accents au degré de proéminence élevé ne peuvent pas se succéder, un groupe de phonation ou un mot prosodique doit être équilibré rythmiquement, etc.

4.1.3. Génération automatique de la mélodie

Trois grands types de modèles peuvent être utilisés pour la génération automatique de la mélodie : le modèle de source vocale, le modèle par points-cibles et les modélisations de contours prosodiques.

Le modèle de source vocale (cf. H. Fujisaki & H. Sudo 1972), inspiré de la réponse d'un filtre de second ordre à des impulsions électriques, repose sur une modélisation de la source vocale. Les contours mélodiques des mots sont organisés en deux composantes : une composante de phrase, liée à la pression subglottique créée par un système musculaire responsable de l'expiration, et une composante d'accent de mot, liée aux variations de la tension des cordes vocales. Développé pour la mélodie du japonais, ce modèle est très économique puisqu'il permet de segmenter l'énoncé en deux types de segments prosodiques uniquement. Si sa pertinence a été prouvée pour le japonais, elle est moins évidente pour des langues comme le français, car un tel modèle ne permet pas de modéliser directement les mouvements montants en fin de groupe caractéristiques de la modalité interrogative. Par ailleurs, un tel système ne peut pas rendre compte de la différence des pentes de transition entre états accentués et non accentués.

Le modèle par points cibles (cf. D. Hirst 1977) consiste à décrire la ligne mélodique sous forme de points stratégiques à hauteur variable : les points cibles, connectés entre eux par des fonctions de transitions.

Les modélisations de contours prosodiques (cf. Ph. Martin 1976 ; J. T'Hart *& al.* 1991) reposent sur la concaténation de segments de droite. Il apparaît ici plus judicieux de considérer un mouvement prosodique dans sa globalité que de déterminer des points raccordables entre eux. Chaque niveau génère un contour fixe, les contours sont ensuites concaténés afin de produire un profil mélodique. La première étape consiste donc à décrire des mouvements mélodiques et à les stocker dans un lexique. Dans un second temps, le développement d'une grammaire permet de définir des règles qui déterminent l'enchaînement de ces mouvements pour former des contours mélodiques.

4.1.4. L'intensité

Contrairement aux phénomènes d'accentuation et de variation mélodique, les variations d'intensité ont souvent été délaissées par les chercheurs en traitement automatique de la parole. En effet, si l'intensité est généralement le premier paramètre qu'un individu associe intuitivement avec l'idée d'accent, cette intuition est trompeuse, du moins en français, où elle joue un rôle très marginal dans les phénomènes d'accentuation, qui sont essentiellement liés aux variations de durée et aux variations mélodiques. Par ailleurs, l'intensité ne peut pas être considérée comme un objet d'étude autonome, dans la mesure où ses variations sont étroitement tributaires de la nature intrinsèque des sons, c'est-à-dire de phénomènes micromélodiques. Néanmoins, des analyses récentes ont montré qu'il était possible de détecter dans le discours des accents d'intensité liés à des phénomènes d'emphase et d'insistance. L'occurrence de ces accents n'est pas très bien définie, il semble cependant qu'elle soit déterminée par des contraintes biologiques et phonotactiques (longueur des mots, contexte accentuel immédiat, etc.). Il est donc

nécessaire de comprendre le fonctionnement de ces accents dans la structure accentuelle globale d'un énoncé et il faudrait en tenir compte dans la génération automatique de la prosodie.

5. Perspectives

Les études prosodiques et les théories linguistiques qui les sous-tendent sont pour beaucoup dans l'amélioration des systèmes de synthèse développés ces dernières années. Elles offrent en effet des outils permettant de structurer la matière sonore en forme linguistique organisée et dotée d'une cohérence interne. Cependant, à la lumière des comptes rendus sur les tests des systèmes de synthèse à partir du texte, force est de constater le caractère encore inhumain, froid et désagréable de la parole synthétique. Les meilleurs systèmes, notamment le système français développé au CNET à Lannion (cf. C. Sorin 1989) sont intelligibles mais manquent fondamentalement de naturel. Si ce handicap n'est pas vraiment un inconvénient pour des tâches spécifiques et limitées — certains même insistent sur la nécessité d'une voix robotique dans des applications précises — il n'est pas acceptable dans le cadre du dialogue homme-machine. Pour les experts, des progrès décisifs ne pourront être obtenus qu'au prix d'une connaissance plus approfondie des mécanismes de production et de perception du langage et notamment des variations suprasegmentales qui sont à l'oeuvre en parole et qui contribuent à produire une parole non seulement intelligible mais également naturelle et agréable, aussi subjectives que puissent être ces notions. En outre, les théories et les modèles développés jusqu'à ce jour sont inopérants pour produire des systèmes de synthèse multistyle et multivoix. En effet, ces théories permettent uniquement de modéliser une stratégie prosodique déterminée liée au style vocal d'un locuteur spécifique. Quant à la question de savoir quelles sont les stratégies possibles en fonction d'un style discursif donné et comment un locuteur varie ses stratégies prosodiques à l'intérieur d'un même énoncé, elle reste ouverte. A ce stade, des connaissances approfondies sur les facteurs ectolinguistiques associés à la parole sont nécessaires.

Enfin, comme nous l'avons déjà souligné, peu de modèles ont été validés dans des tâches de reconnaissance, il semble en effet que les indices prosodiques proposés soient insuffisamment fiables pour être utilisés de façon pertinente dans de telles tâches. Pour les experts, l'absence d'une théorie unifiée, capable de rendre compte de l'ensemble des phénomènes prosodiques, semble être une des raisons principales de ces limitations (cf. A. Waibel 1990).

Anne LACHERET-DUJOUR
(Université de Caen, LIMSI-CNRS et ELSAP-CNRS)

Repères bibliographiques

1. Prosodie et théories linguistiques :

CAELEN-HAUMONT, G. (1991) : *Stratégies des locuteurs en réponse à des consignes de lecture d'un texte : analyse des interactions entre modèles syntaxiques, sémantiques, pragmatiques et paramètres prosodiques,* Thèse de Doctorat d'Etat, Institut de la Communication Parlée, Grenoble.

[Présentation d'un modèle énonciatif qui repose sur la prise en compte d'une hiérarchisation des unités intonatives en français.]

DELATTRE, P. (1966) : Les dix intonations de base du français, *French Review*, 40, Illinois, American Association of Teachers of French, 1-14.

[Proposition d'une modélisation de l'intonation des phrases françaises.]

DELATTRE, P. (1967) : La nuance de sens par l'intonation, *French Review*, 41, Illinois, American Association of Teachers of French, 326-339.

[Etude portant sur la congruence entre structures grammaticales françaises et variations intonatives.]

DELL, F. *& al.* (1984) : *Forme sonore du langage, Structure des représentations en phonologie,* Paris, Hermann.

[Présentation des développements de la phonologie générative dans une perspective métrique et autosegmentale.]

DI CRISTO, A. (1975) : Recherches sur la structuration prosodique de la phrase française, *6èmes Journées d'étude sur la parole*, Toulouse, Société Française d'Acoustique, 95-116.

[Proposition d'un modèle phono-syntaxique applicable à la reconnaissance et fondé sur l'étude acoustique et perceptuelle des structures prosodiques du français.]

DUCROT, O. TODOROV, T. (1972) : *Dictionnaire encyclopédique des sciences du langage*, Paris, Seuil.

[Ouvrage d'introduction aux concepts fondateurs de la linguistique contemporaine.]

HIRST, D. (1986) : Représentation phonologique et phonétique des langues naturelles : présentation d'un projet, *Bulletin de l'Institut de Phonétique d'Aix*, 10, Aix-en-Provence, 127-149.

[Tentative de modélisation des paramètres de représentation phonologique et phonétique de l'intonation des langues naturelles.]

MARTIN, Ph. (1973) : Les problèmes de l'intonation : recherches et applications, *Langue française*, 19, Paris, Larousse, 4-32.

[Approche théorique de l'intonation dans une perspective fonctionnelle appliquée à l'enseignement du français.]

MARTIN, Ph. (1987) : Structure rythmique de la phrase française, statut théorique et données expérimentales, *16e Journées d'étude sur la parole*, Hammamet, Société Française d'Acoustique, 255-258.

[Présentation d'un modèle pour l'analyse de la structure rythmique des phrases françaises.]

ROSSI, M. *& al.* (1981) : *L'intonation, de l'acoustique à la sémantique*, Paris, Klincksieck, Coll. "Etudes Linguistiques", 15.

[Ouvrage de référence sur les recherches appliquées à l'intonation du français dans une perspective perceptive, acoustique et linguistique.]

2. Prosodie et traitement automatique de la parole :

CALLIOPE (1989) : *La parole et son traitement automatique*, Paris, Masson.

[(cf. *supra*, chapitre 1)].

• Prosodie et synthèse de la parole :

BAILLY, G. (1986) : Un modèle de congruence relationnel pour la syntaxe de la prosodie du français, *15èmes Journées d'étude sur la parole*, Aix-en-Provence, Société Française d'Acoustique, 75-78.

[Présentation d'un modèle prosodique pluriparamétrique pour la synthèse à partir du texte.]

CHARPENTIER, F. & MOULINES, E.(1989) : Nouvelles techniques de synthèse de la parole, *L'Echo des recherches* , 137, 37-46.

[Article de vulgarisation où sont présentés les résultats obtenus ces dernières années au CNET à Lannion en synthèse de la parole.]

CHOPPY, C., LIENARD, J.S. (1977) : Prosodie automatique pour la synthèse par diphonèmes, *8èmes Journées d'étude sur la parole*, Aix-en-Provence, Société Française d'Acoustique, 211-217.

[Présentation d'un algorithme de génération automatique de la prosodie sans analyse syntaxique utilisé en synthèse à partir du texte.]

DANLOS, L. *& al.* (1985) : Synthèse de messages oraux à partir d'une représentation sémantique, *14e Journées d'Etude sur la parole*, Paris, Société Française d'Acoustique, 118-121.

[Article présentant une application de la génération automatique de texte à la synthèse de la parole en français.]

FUJISAKI, H. & SUDO, H. (1972) : A Generative Model for the Prosody of Connected Speech in Japanese, *IEEE International Conference on Acoustic Speech and Signal Processing*, Newton, Institute of Electrical and Electronics Engineers Signal Processing Society, 140-143.

[Article présentant le développement d'un modèle de génération automatique de la mélodie pour le japonais.]

T'HART, J. & AL. (1991) : *A Perceptual Study of Intonation : an Experimental Phonetic Approach to Speech Melody,* Cambridge, Cambridge University Press.

[Présentation d'une méthodologie d'analyse des faits prosodiques fondée sur une approche acoustique et perceptive.]

HIRST, D. (1977) : *Intonative Features,* The Hague, Mouton, Coll. "Janua Linguarum".

[Ouvrage de référence en phonologie et en phonétique. Une des trois grandes méthodologies existantes pour la dérivation automatique d'une structure prosodique hiérarchisée et linguistiquement pertinente, utilisable en traitement automatique de la parole, y est exposée.]

LARREUR, D. & EMERARD, F. (1977) : Analyse de la structure intonative de la phrase française, *8èmes Journées d'Etude sur la Parole*, Aix-en-Provence, Société Française d'Acoustique, 227-236.

[Présentation d'un ensemble de règles prosodiques applicables à la synthèse à partir du texte en français.]

MARTIN, Ph. (1976) : Perception des séquences de contours prosodiques de phrases synthétisées, *9° journées d'étude sur la parole*, Lannion, Société Française d'Acoustique, 21-29.

[Synthèse de contours mélodiques et structure syntaxique : évaluation perceptive.]

SORIN, C. (1989) : Speech Synthesis Applications and Research : Present and Future, *Congrès International d'Acoustique*, 1, Belgrade, Sava Centar, 29-40.

[Présentation synthétique et claire de l'état de l'art et des principaux développements récents en synthèse de la parole à partir du texte.]

VAISSIERE, J. (1975) : Caractérisation des variations de la fréquence du fondamental dans les phrases françaises, *6èmes Journées d'étude sur la parole,* Toulouse, Société Française d'Acoustique, 39-50.

[Etude portant sur les rapports entre la structure grammaticale de la phrase française et les variations mélodiques.]

• Prosodie et reconnaissance de la parole :

WAIBEL, A. (1990) : *Prosody and Speech Recognition,* Londres, Pitman.

[Etude de synthèse présentant l'utilisation de la prosodie en reconnaissance de la parole et décrivant avec un regard critique l'état de l'art et l'avancement des recherches dans ce domaine.]

• **Autres :**

FRAISSE, P. (1967) : *Psychologie du temps,* Paris, PUF.

[Dans le cadre des méthodes proposées par la psychologie du comportement, et dans une perspective de renouvellement des analyses classiques, l'auteur présente un ensemble de réflexions sur les différentes manières dont l'homme s'apdapte aux conditions temporelles de son existence. Il essaie de répondre aux questions suivantes : comment l'homme se conditionne-t-il au temps, comment le maîtrise-t-il, enfin comment le perçoit-il ?]

3
MORPHOLOGIE

En traitement automatique, l'analyse morphologique consiste à segmenter un texte en **unités élémentaires** auxquelles sont attachées des connaissances dans le système : une fois cette segmentation effectuée, ce n'est plus le texte qui est manipulé, mais une liste ordonnée de telles unités (ou plusieurs listes, en cas d'ambiguïtés, réelles ou "artificielles").

Concrètement, pour le traitement d'un texte **écrit**, on part d'une chaîne de caractères typographiques (saisie sur ordinateur d'une façon ou d'une autre) et on essaie de la découper de façon à ce que chaque segment corresponde à une unité répertoriée dans le système.

Prenons un exemple : la segmentation de la chaîne de caractères *Jean a mangé des pommes de terre* pourra aboutir à la suite d'unités (u_1, u_2, u_3, u_4, u_5), dans laquelle chaque u_i correspond à une unité répertoriée (comme on le verra tout au long de ce chapitre, le nombre et la nature de ces unités dépendent d'un certain nombre de choix théoriques et pratiques). Dans le système, toutes sortes d'informations peuvent être associées aux u_i, comme par exemple :

u_1 : segment : *Jean*

informations morpho-syntaxiques : nom propre, masculin, singulier
informations sémantiques : animé humain, prénom ...

u_2 : segment : *a mangé*

forme lemmatisée : *manger*
informations morpho-syntaxiques : verbe, passé composé, indicatif, 3e pers., singulier, constructions : transitif, ...

informations sémantiques : rôles actanciels associés : (agent : animé, objet : nourriture), activité, sorte d'absorption, ...

u_3 : segment : *des*

forme lemmatisée : *un*
informations morpho-syntaxiques : déterminant, pluriel, masculin ou féminin
informations sémantiques : marque d'extraction d'un nombre indéterminé d'occurrences discrètes, ...

u_4 : segment : *pommes de terre*

forme lemmatisée : *pomme de terre*
informations morpho-syntaxiques : nom commun, féminin, pluriel
informations sémantiques : entité discrète, inanimé, sorte de nourriture, ...

u_5 : segment : .

informations morpho-syntaxiques : ponctuation, délimiteur de phrase, ...

Dans le cas du traitement d'un texte **oral**, le but à atteindre est approximativement le même, mais la tâche est plus complexe puisque le point de départ est une chaîne sonore dans laquelle il faut d'abord (ou parallèlement) reconnaître des phonèmes (cf. *supra* chapitre 1). Un problème analogue se pose avec les écritures manuscrites, où la reconnaissance des caractères est loin d'aller de soi. Rappelons par ailleurs que la morphologie de l'**écrit** est, sur bien des points, distincte de la morphologie de l'**oral** correspondante. Pour n'en prendre qu'un exemple, la phrase *Leurs livres traînaient sur leurs bureaux* comporte à l'écrit cinq marques (désinences) de pluriel (*-s, -s, -ent, -s, -x*) alors qu'elle n'en comporte aucune à l'oral (phoniquement, la phrase ne se distingue pas de *Leur livre traînait sur leur bureau*, ni de *Leurs livres traînaient sur leur bureau* ou de *Leur livre traînait sur leurs bureaux*).

Pour réaliser un analyseur morphologique, il faut effectuer une série de **choix**, qui engagent tout à la fois théorisation linguistique et réalisation informatique ; ces choix concernent les types d'**unités** retenues (cf. *infra*, § 1.), les types d'**informations** associées à ces unités (cf. *infra*, § 2.), et enfin les **méthodes de reconnaissance** des unités (cf. *infra*, § 3.).

1. Les unités

Comme on va le voir, les difficultés théoriques et pratiques rencontrées dans le choix des unités tournent autour de la définition et du statut de ce que l'on appelle un **mot**.

1.1. Le mot : un choix "naturel"

A première vue, le choix des unités peut sembler simple : on dira que l'analyse morphologique découpe le texte en **mots** ; le mot, tel qu'il s'est imposé dans la forme écrite de la langue, étant défini comme une chaîne de caractères comprise entre deux blancs ou entre un blanc et un signe de ponctuation.

Remarquons tout de suite que cette définition doit être légèrement modulée, même si l'on ne la remet pas en cause sur le fond, en raison du rôle particulier de deux signes de ponctuation : l'apostrophe et le tiret.

Considérons tout d'abord l'**apostrophe**. On voudra sans doute considérer *aujourd'hui* ou *d'abord* comme un seul mot, alors que *l'arbre* ou *j'arrive* doivent être segmentés en deux mots. La liste des mots tels que *aujourd'hui* étant très restreinte et fermée, il suffit de l'établir une fois pour toutes, et de rectifier la définition initiale en explicitant cette liste d'"exceptions".

Considérons à présent le **tiret**. Le problème est analogue, mais plus difficile. Pour traiter des séquences comme : *porte-monnaie*, *c'est-à-dire*, *avant-hier*, *dix-sept*, *cet homme-là*, *lui-même*, *voulez-vous*, etc., il faut pouvoir distinguer les cas où le tiret doit être considéré comme interne à une unité (on pourra alors décider de supprimer le tiret, lors d'un pré-traitement morphographique du texte), et ceux où il joue un rôle spécifique dans une construction syntaxique. Contrairement à l'apostrophe, la liste des termes lexicaux comportant un tiret est vaste et ouverte : la règle doit donc plutôt expliciter les constructions syntaxiques, en nombre limité, où le tiret apparaît. (Une attention particulière doit être portée aux cas où ces constructions s'accompagnent de modifications morphologiques qui dépassent l'apparition du tiret lui-même : *puissé-je*, *arrive-t-il*, ...).

Mais, comme nous allons le voir, au-delà de ces premières difficultés, une série de problèmes, plus ou moins difficiles à résoudre, peuvent faire douter de l'opérationnalité du choix "naturel" dont nous sommes partis.

1.2. Le mot : un choix problématique

Le choix des unités doit obéir à deux impératifs : la segmentation ne doit pas être trop difficile à effectuer, et les unités doivent être suffisamment cohérentes et "significatives" pour faciliter les traitements ultérieurs. Cette

double contrainte se heurte à toute une série de phénomènes répertoriés de longue date par les linguistes : élisions, amalgames, flexions, dérivations, compositions, etc., qui conduisent à s'interroger sur la notion de "mot", telle qu'elle vient d'être présentée. Les lignes qui suivent illustrent certaines des difficultés de segmentation rencontrées.

Dans les cas d'**élision**, doit-on considérer comme une unité à part entière un terme élidé, ou doit-on rétablir la forme pleine ? La restitution de la forme pleine, facile dans certains cas (comme par exemple *j'*, *m'*, *qu'*, ...), est plus délicate dans d'autres (ainsi *d'* peut provenir de *de* ou de *des*, *l'* peut provenir de *le*, de *la* ou ne pas être le résultat d'une élision, comme dans *l'on sait que*, etc.). Ainsi des connaissances sémantico-pragmatiques sont-elles nécessaires pour rétablir *la* dans *il vit la bestiole dans le trou et il l' (= la) enferma ;* dans un cas de ce genre, le choix au niveau de l'analyse morphologique consiste donc soit à conserver une unité *l* de genre indéterminé, soit à introduire une ambiguïté artificielle entre *le* et *la* qui ne pourra être levée que plus tard. De plus, il faut aussi veiller à ne pas perdre d'information en rétablissant la forme pleine : ainsi *c'* provient de *ce*, bien sûr, mais est forcément un pronom sujet, alors que *ce* est très polyvalent.

Certains "mots", au sens du paragraphe précédent, résultent, on le sait, de l'**amalgame** de deux unités existantes. Faut-il alors rétablir les deux unités (qui jouent chacune un rôle syntaxique spécifique) pour faciliter l'écriture des règles ultérieures ? Le gain semble évident pour des amalgames tels que *au*, *aux*, *auquel*, etc. C'est déjà un peu plus coûteux pour *du*, *des*, etc. où l'on introduit forcément à ce stade l'ambiguïté artificielle entre un simple déterminant (ex : *du* partitif) et un amalgame (*du* = préposition *de* + déterminant *le*). Jusqu'où doit-on étendre ce mécanisme ? Faut-il considérer comme un amalgame l'adverbe interrogatif *pourquoi*, ou même, en poussant à l'extrême, décomposer le pronom relatif *qui* en un amalgame d'une marque de relatif et d'un pronom clitique sujet (dont le genre, le nombre et la personne seraient indéterminés), ou encore la conjonction de subordination *quand* en un amalgame d'un adverbe de temps et de la conjonction *que* ?

Par ailleurs, on sait aussi que nombre de "mots" connaissent des variations de forme, à savoir des **flexions.** Les phénomènes de flexion occupent une place centrale dans l'analyse morphologique, du fait de leur importance quantitative (en français il s'agit essentiellement de la flexion verbale) et de leur relative facilité de traitement due au caractère fermé de l'inventaire des flexions dans une langue : à tel point que souvent l'on assimile (abusivement) analyse morphologique et traitement des formes fléchies.

Nous verrons au § 3.1. que le traitement algorithmique des flexions n'est pas un problème difficile. La question qui nous intéresse ici est celle du **statut** des désinences : dans la plupart des systèmes elles sont traitées comme une série d'attributs (de temps, de mode, de nombre, etc.) qui sont ajoutés à la forme dite "lemmatisée" (forme conventionnellement choisie comme entrée

de dictionnaire) ; par exemple la forme *chanterions* est traitée comme la forme lemmatisée *chanter* suivie d'un attribut "mode" (valeur "conditionnel"), d'un attribut "temps" (valeur "présent"), d'un attribut "personne" (valeur "1é") et d'un attribut "nombre" (valeur "pluriel"). Mais on pourrait aussi les considérer comme des unités morphologiques à part entière (à la manière des "morphèmes" de la linguistique), et traiter la forme fléchie comme un amalgame de plusieurs unités, autonomes ou non, mais susceptibles d'apparaître indépendamment les unes des autres ; ainsi *chanterions* serait analysée comme constituée des quatre unités *chant-* , *-er-*, *-i-* et *-ons* (l'unité *-i-* se retrouvant par exemple dans *chantions*, et l'unité *-ons* dans *chantons*).

Autre enjeu théorique, rarement évoqué : jusqu'où étendre le phénomène de flexion ? Ainsi, en français, doit-on admettre (et traiter comme tels) des phénomènes de déclinaisons et, par exemple, représenter *je*, *me*, *mon*, *moi*, *la mienne*, etc., comme les formes déclinées d'un même pronom personnel MOI (*je* = MOI + marque de nominatif, *mon* = dét. *le* + MOI + marque de génitif) ? Un traitement similaire pourrait être envisagé pour les pronoms relatifs.

Si la prise en compte des flexions est relativement aisée dans un analyseur morphologique, en revanche celle des **dérivations** l'est beaucoup moins : les mots dérivés constituent en effet, comme d'ailleurs les mots composés, un inventaire ouvert, en constante évolution ; de plus, leur analyse pose des problèmes théoriques difficiles. Certes il existe des dérivations extrêmement productives en français, comme la dérivation d'un adverbe à partir d'un adjectif suivi du suffixe *-ment* ou la préfixation d'un verbe par la forme *re-* : la représentation de ces découpages fournit un gain appréciable pour la description syntaxique et sémantique des unités dérivées (même si un préfixe comme *re-* est lui-même polysémique, son apport sémantique à un verbe donné obéit à des régularités qui peuvent être explicitées). Mais là encore, il est difficile de savoir jusqu'où systématiser ces mécanismes. Les nombreux "trous" et irrégularités apparentes du système de dérivation peuvent faire douter de l'efficacité d'une prise en compte des faits de dérivation dans un analyseur morphologique. Pour n'en citer que quelques exemples : pourquoi existe-t-il les deux dérivés *décentrage* et *décentrement* dans le cas du terme préfixé, mais seulement *centrage* et pas * *centrement* dans le cas du terme non-préfixé ? si *carpette* (au sens de "tapis") est à considérer comme une forme complexe, faut-il dériver celle-ci d'une base **carp-* inexistante (aucun rapport étymologique ou sémantique avec *carpe* "poisson") ? pourquoi *truchement* n'est-il pas décomposable en *truch-ement*, comme *errement* ? pourquoi un *anti-clérical* est-il un mot dérivé signifiant "opposé au clergé", alors que l'*antimoine* ne signifie pas "opposé au moine" et ne doit pas plus être décomposé que l'*antilope* ? Sur les dérivations en français, on pourra consulter D. Corbin (1987) et D. Corbin (éd.) (1991).

Toutes les difficultés que nous venons d'évoquer concernent des "mots" (au sens typographique évoqué au début de ce chapitre) recouvrant en fait

plusieurs unités. Comme on va le voir à présent, l'inverse se rencontre également : à savoir l'existence d'une unité unique distribuée sur une séquence de plusieurs "mots" typographiques.

C'est le cas, tout d'abord, des **unités discontinues,** comme les négations (*ne...pas, ne...plus*) à l'intérieur desquelles s'intercale le verbe ou l'auxiliaire conjugué (ex : *il ne mange pas, il n'a plus mangé*), sauf devant un infinitif (ex : *il a décidé de ne pas manger*). Toutefois, il est clair que la décision de traiter de telles séquences comme des unités uniques discontinues engage des choix théoriques : même sans évoquer l'étymologie de ces négations (dont le second membre était, clairement, une partie du discours indépendante et autonome : substantif *pas*, adverbe *plus*), il reste que la frontière est difficile à tracer : doit-on traiter *ne jamais* comme une unité discontinue, alors que *jamais* se rencontre en emploi autonome (ex : *Si jamais tu viens...*) ? Autre exemple, les temps composés : des chaînes de longueur considérable peuvent s'insérer entre l'auxiliaire et le participe (*Il a, sans le faire exprès, avec toute la candeur qui fait son charme, tout fait capoter*), et, pour les verbes utilisant l'auxiliaire *être*, la frontière avec la construction "*être* + adj." pose des problèmes délicats (comparez *Il est mort depuis dix ans* et *Il est mort il y a dix ans*).

Le même ordre de difficulté se retrouve dans le cas des **locutions**, des **formes figées** et des **mots composés**, qui font toucher du doigt l'inadéquation de la définition du mot dont nous sommes partis. Certains "mots" n'ont d'existence que dans une locution (*tandis* n'apparaît que dans *tandis que*) ; d'autres peuvent tantôt entrer dans une locution tantôt rester autonomes (c'est le cas de *alors que*, comme le montre l'ambiguïté de *je pensais alors que je devais agir*). Cette même dualité se rencontre également avec les mots composés, susceptibles de prendre une spécificité sémantique difficilement calculable par composition (*pomme de terre* n'est pas parfaitement défini — c'est le moins qu'on puisse dire — par la glose *chose en forme de pomme qui croît sous terre*). On a affaire à un continuum qui va d'expressions si soudées que l'absence de tiret semble arbitraire, jusqu'à des expressions qui peuvent être fort complexes (*prendre la poudre d'escampette,...*) et / ou accepter des insertions de toutes sortes : le processus de lexicalisation connaît toutes sortes de degrés. G. Gross (1990, p.58) comptabilise ainsi jusqu'à 512 "degrés de figement" possibles, pour la seule construction de nom composé en "nom + adjectif", à partir de neuf critères linguistiques. Pour rendre compte de ce continuum de lexicalisation à l'oeuvre dans les mots composés, l'approche connexionnniste paraît bien adaptée (cf. par exemple Ch. Jacquemin & P. Cadiot 1992).

L'existence d'un tel continuum oblige à choisir où l'on s'arrête dans la définition de l'unité morphologique. Plusieurs réponses sont possibles. La position minimale consiste à n'admettre comme unités que les séquences dont l'un des membres n'est jamais autonome (comme *tandis que*). La définition est simple, mais malheureusement elle ne couvre qu'une partie infime du phénomène. Une autre position consiste à considérer comme unités les

séquences qui sont insécables : ainsi *encore que*, *au cours de*, etc., aussi bien que *pomme de terre*, *chemin de fer*, etc., seront considérées comme des unités, tandis que *au moment où* par exemple ne le sera pas (on peut dire *au moment précis où*). Cette solution force bien sûr à introduire des ambiguïtés artificielles, puisque l'on ne peut préjuger à ce stade de la présence effective du mot composé ; ainsi dans : *Il a couvert sa pomme de terre, pomme de terre* peut être un mot composé ou non ; mais dans *Il est plus fort encore que je ne le pensais*, on n'a pas affaire à la locution conjonctive *encore que*. Des positions maximalistes sont également possibles, mais il faut alors faire conjointement l'analyse morphologique et l'analyse syntaxique : en particulier, le regroupement de l'auxiliaire et du participe passé, considéré comme un constituant "temps composé" discontinu, ne peut se faire qu'après avoir identifié les syntagmes insérés entre eux.

1.3. Le mot ou le morphème ?

Comme on le voit, la définition "spontanée" de l'unité morphologique s'avère moins simple qu'il n'y paraît. Toute une série de choix sont nécessaires, et le danger essentiel, c'est de perdre, dans des choix "au coup par coup", la cohérence qui seule peut permettre un développement du système pour une large couverture. La question se pose donc de savoir sur quelles bases théoriques il est possible d'effectuer ces choix.

On sait que les difficultés qui viennent d'être rappelées ont été épinglées depuis longtemps en linguistique : c'est précisément pour tenter d'y remédier que la tradition structuraliste avait proposé de bannir le mot, et de travailler avec le **morphème**, défini comme l'unité minimale significative (cf. notamment A. Martinet 1960, rééd. 1991). Néanmoins (si l'on fait exception des exercices bien ciblés forgés dans des langues aux formations morphématiques très régulières : swahili, etc.) ces tentatives se sont elles aussi heurtées à de redoutables problèmes de découpage : il n'existe pas de méthode univoque permettant de découper un texte français en morphèmes !

En définitive, la difficulté d'élaborer une théorie opératoire des unités morphologiques tient à la nature même des langues. La fiction d'une définition bien nette se heurte à la réalité d'un continuum de situations : toute définition classera correctement des situations prototypiques mais aura du mal avec les frontières, que l'on fait entrer "de force", avec une part irréductible d'arbitraire, dans le cadre théorique. Notons dès à présent que cet état de fait se retrouve à tous les niveaux du traitement de la langue.

1.4. Le mot : un compromis

Des considérations précédentes il ressort que le mot, tel qu'il s'est imposé dans la langue écrite, ne constitue pas un point de départ plus mau-

vais qu'un autre. Ce à quoi vont s'attacher les systèmes pour aménager ce point de départ (et donc choisir entre les diverses possibilités évoquées au §1.2), c'est à établir le meilleur compromis possible entre la cohérence des choix et la facilité de réalisation de l'analyse : éviter l'introduction de trop d'ambiguïtés artificielles pour ne pas risquer l'explosion combinatoire dans la suite des traitements, résoudre le maximum de problèmes de morphologie avec des connaissances limitées tout en n'hésitant pas à laisser en suspens ceux qui exigent une analyse syntaxique complète, essayer d'obtenir la représentation morphologique la mieux adaptée au traitement syntaxique choisi, pour que la formulation des règles syntaxiques soit la plus régulière possible. Ainsi on n'hésitera pas à ignorer les temps composés dans un premier temps, quitte, à un autre stade de l'analyse, à revenir sur la représentation morphologique pour récupérer le verbe véritable avec ses temps et modes. Il peut en être de même pour les formes verbales figées (*faire fi*, *prendre conscience*, etc.) ou les négations. A l'opposé, on traitera dès le départ comme unités des expressions figées parfois plus complexes, mais qui posent moins de problèmes et dont l'analyse syntaxique serait difficile à décrire (ex. *à qui mieux mieux* sera immédiatement traité comme une unité adverbiale, d'autant plus facilement que l'analyse syntaxique de cette expression n'est pas évidente, de même que *comme qui dirait* ou *à la va comme je te pousse*).

2. Les informations associées aux unités

Dans le dictionnaire, les informations associées aux unités peuvent être de nature diverse. Au plan morphologique, elles concernent principalement la catégorie et les flexions de l'unité.

2.1. La catégorie

En matière de choix de catégories, signalons d'emblée qu'il n'existe pas de consensus : le nombre et la nature des catégories retenues varient d'un analyseur à l'autre. Or le **choix des catégories** est essentiel : ce sont en effet les latitudes combinatoires de ces catégories que les règles grammaticales ont pour rôle de caractériser. On a donc intérêt à avoir les catégories les plus spécifiques possibles, de manière à décrire leur comportement syntaxique de façon précise. Ainsi une catégorie des pronoms clitiques compléments aurait l'avantage de permettre l'écriture simple des règles qui régissent leur position antéposée par rapport au verbe, alors qu'une catégorie très générale des pronoms conduirait à compliquer les règles : une règle admettant la possibilité que n'importe quel pronom puisse être complément du verbe devant lequel il se trouve, et qu'à l'inverse n'importe quel pronom puisse se situer derrière le verbe, conduit à multiplier les ambiguïtés artificielles, d'autant plus que cer-

tains clitiques (*le, la, les, leur, ...*) sont portés par des unités polycatégorielles parmi les plus fréquentes. Si l'on ne dispose que d'une catégorie pour les pronoms, toutes les règles les concernant devront donc préciser le type de pronom par un autre moyen (sous-catégorie, ou attribut spécifique), ce qui compliquera d'autant ces règles.

A l'inverse, la multiplication des catégories peut entraîner une augmentation des ambiguïtés artificielles en multipliant les cas de mots polycatégoriels (qui sont de toute façon très nombreux, comme on va le voir). Il faut donc là aussi chercher à réaliser le meilleur compromis possible.

Le recours à la traditionnelle classification en **parties du discours** (noms, adjectifs, verbes, adverbes, articles, pronoms, conjonctions, prépositions, interjections) pose un certain nombre de problèmes, dont nous allons donner quelques exemples.

Les phénomènes de **polycatégorie** de toutes sortes sont massifs. On insiste souvent sur les cas d'homonymie catégorielle (ainsi par exemple la séquence *ferme* recouvre-t-elle en synchronie, malgré une étymologie commune, trois homographes : le substantif désignant l'"exploitation agricole", l'adjectif renvoyant à quelqu'un ou à quelque chose de "solide", et le verbe conjugué signifiant "clore"). Mais les cas vraiment importants sont ici ceux où la polycatégorie provient d'une même unité prenant des significations déductibles les unes des autres dans des emplois syntaxiques différents (ce que l'on appelle parfois "dérivation impropre") : ainsi entre nom et adjectif (de *petit, rouge, juste* à *informatique* et *linguistique*), entre nom et verbe à l'infinitif (*rire, pouvoir, avoir, manger*, etc.), entre adjectif et adverbe (*clair, fort, juste*). La question se pose de savoir si l'on n'a pas, dans certains cas, intérêt à créer une catégorie "intermédiaire" (une classe des "adjectifs-noms" ou des "adjectifs-adverbes" par exemple) plutôt que d'avoir des doubles entrées systématiques dans le dictionnaire ; ainsi l'unité *nouvelle* peut-elle être d'emblée catégorisée comme "adjectif-nom", quitte à préciser au niveau des règles grammaticales que cette catégorie intermédiaire admet tout à la fois les latitudes combinatoires de l'adjectif (comme dans *Elle était toute nouvelle*) et celles du nom (comme dans *La nouvelle vint se présenter dès le lendemain*). Notons cependant que, dans ce dernier exemple, cela n'empêchera pas d'avoir deux entrées distinctes : *nouveau* "adjectif-nom" et *nouvelle* "nom" (par exemple dans *La nouvelle fut immédiatement diffusée à la radio*).

Certaines formes verbales ont des constructions très différentes des formes conjuguées standard : l'infinitif et les participes. Ainsi, pour ne prendre qu'un exemple, un infinitif peut faire partie d'un groupe prépositionnel, là où la forme conjuguée réclame une conjonctive (*Le fait de me marier* à opposer à *Le fait que tu te maries*). Ne faut-il pas, pour alléger l'écriture des règles syntaxiques en faire des catégories à part ? Dans le même ordre d'idées, le verbe *être*, dont les constructions diffèrent assez radicalement des autres verbes, ne mérite-t-il pas une catégorie à lui tout seul ? La catégorie

des conjonctions est manifestement trop large. Ne faut-il pas au minimum distinguer les conjonctions de coordination des autres ? Une analyse de la classe des adverbes fera apparaître aussi plusieurs sous-classes (par exemple les comparatifs). Nous avons déjà parlé des pronoms clitiques. La même remarque vaut pour les pronoms relatifs. Faut-il distinguer aussi dès ce niveau les pronoms interrogatifs ? Combien d'entrées catégorielles distinctes pour une unité comme *que* ou *où* ?

Comme on le voit, bien des choix, qui s'expriment dès le niveau de la représentation morphologique, vont conditionner la suite des traitements. Là encore, une abondante littérature linguistique témoigne des efforts accomplis pour proposer des classifications plus cohérentes et opérationnelles. Si ces travaux permettent de mieux comprendre les problèmes et de guider les choix, il n'en reste pas moins qu'aucune classification réellement satisfaisante ne s'est imposée : on se heurte au même problème d'un continuum rebelle à toute classification rigide. Les compromis entre cohérence théorique et efficacité opérationnelle restent donc inévitables.

Dans le cas où le mot n'existe pas en tant que tel dans le dictionnaire, et où il a été reconnu par application de règles de composition ou de dérivation, la catégorie doit aussi être elle-même calculée. Pour assigner une catégorie à un mot **dérivé**, certaines règles linguistiques sont applicables. Ainsi, en français, alors que les mots **préfixés** conservent en général la catégorie de la base, qui est souvent un mot susceptible d'un emploi autonome, parfois lui-même déjà suffixé (*endormir* est un verbe comme *dormir* ; *inintéressant* un adjectif comme *intéressant* ; *prévision* un substantif comme *vision*), en revanche la catégorie des mots **suffixés** est donnée par le suffixe. Des études menées dans l'équipe de M. Gross dès les années 70 ont permis d'établir des listes de suffixes, tels que :

- suffixes nominaux dérivant des noms à partir de bases verbales (comme *trembl-ement, priv-ation*), ou à partir de bases adjectivales (comme *petit-esse, grand-eur, futil-ité, ferme-té*)

- suffixes adjectivaux dérivant des adjectifs à partir de bases nominales (comme *enfant-in*) ou à partir de bases verbales (comme *cass-able*)

- suffixes verbaux dérivant des verbes à partir de bases nominales (comme *fleur-ir*) ou à partir de bases adjectivales (comme *roug-ir*, *minim-iser*)

- suffixes adverbiaux dérivant des adverbes à partir de bases adjectivales (comme *terrible-ment*).

L'intérêt de telles études pour un traitement automatique est, entre autres, de permettre d'identifier, à partir de l'affixe, la catégorie d'un mot nouveau inconnu, dérivé selon les règles de la langue.

En ce qui concerne la catégorie des mots **composés**, les travaux de l'équipe de M. Gross (cf. en particulier les articles consacrés aux mots composés dans Bl. Courtois & M. Silberztein (eds.) 1990) ont montré qu'en fran-

çais la constitution des mots composés de catégorie nominale obéit à une grande diversité de schémas ; exemples de tels schémas : "N N" (substantif substantif) comme *bateau mouche*, "V N" (verbe substantif) comme *porte-plume*, "N A" (substantif adjectif) comme *chaise longue*, "A N" (adjectif substantif) comme *long métrage*, "N *de* N" (substantif *de* substantif : classe numériquement la plus importante) comme *pomme de terre*, "N *à* N" (substantif *à* substantif) comme *pâte à crêpes*, "N *à* V" (substantif *à* verbe) comme *machine à laver*, etc. Là encore, un traitement automatique peut, en s'appuyant sur de tels recensements, calculer la catégorie d'un mot composé inconnu à partir des catégories de ses constituants.

2.2. Les informations flexionnelles

En général les informations flexionnelles sont ajoutées à l'unité qui les porte sous forme de **valeurs d'attributs** (encore que, comme on l'a vu au § 1.2. *supra*, elles puissent aussi être traitées comme des unités à part entière) de genre, de nombre, de personne, de temps, etc., ou même de fonction syntaxique (pour certains pronoms, si ce choix a été fait). Outre leur importance propre au plan sémantique, ces informations jouent un rôle dans l'analyse syntaxique en raison des phénomènes d'accord.

Un point important est de permettre aux attributs de rester **indéterminés** ou **multivalués** afin d'éviter d'engendrer des ambiguïtés artificielles de façon massive : il est plus efficace (encore que cela dépende de la forme des règles syntaxiques) de traiter *les*, *des* ou *ces* comme des unités dont le genre est indéterminé, plutôt que de donner chaque fois l'alternative entre une unité avec l'attribut "genre masculin" et une autre avec l'attribut "genre féminin".

2.3. Les autres informations

Les autres informations attachées aux unités dans le dictionnaire concernent les constructions syntaxiques et les caractérisations sémantiques ; elles dépassent le cadre de la morphologie et seront traitées dans les chapitres suivants. Il convient toutefois d'évoquer dès ici la question de la **polysémie** : combien d'entrées lexicales pour une unité ayant plusieurs sens ? La tradition linguistique, se fondant sur l'unicité du signe, pousse à considérer que dans le cas de la polysémie il n'y a qu'une seule unité lexicale (et donc une seule entrée), contrairement aux cas de collision homonymique (comme *avocat* "fruit" ou "plaideur", ou *bière* "cercueil" ou "boisson") ; en revanche les traitements automatiques tendent à l'**atomisation** : ils pratiquent de façon assez systématique des "dégroupements homonymiques" chaque fois qu'à une unité peuvent être associés plusieurs comportements syntaxiques et / ou sémantiques différents (c'est notamment ce qui se pratique le plus couramment dans le domaine de la lexicographie automatisée). Le risque est évidemment de multiplier indéfiniment le nombre d'entrées lexicales (attention à

l'explosion combinatoire) et de perdre ce qui, derrière les différences, fonde l'unicité de l'expression.

3. Les méthodes de reconnaissance des unités

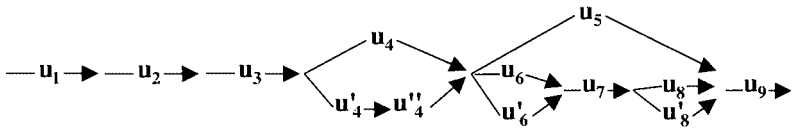
Au delà de la diversité des choix théoriques possibles (même si ceux-ci ne sont pas toujours explicités), la tâche informatique de segmentation du texte en unités morphologiques nécessite la mise en place d'algorithmes qui obéissent *grosso modo* au même principe.

3.1. Le principe de la reconnaissance en morphologie

Le principe de la reconnaissance est relativement simple, du moins pour les textes écrits typographiés : comme on l'a vu, il s'agit de constituer une liste d'unités répertoriées dans le système, chaque unité représentant une partie du texte traité. Pour ce faire, l'analyseur doit comparer tour à tour les portions de texte, constituées d'un ou plusieurs mots, avec les chaînes de caractères stockées dans le dictionnaire (ou les dictionnaires, s'il y en a plus d'un). Ainsi, pour reprendre l'exemple du début de ce chapitre (*Il a mangé des pommes de terre.*), les chaînes *Il, Il a, Il a mangé, Il a mangé des,* etc. puis *a, a mangé, a mangé des,* etc. doivent théoriquement être examinées (en fait, bien entendu, toutes ces chaînes ne sont pas réellement traitées : s'il n'y a pas d'entrée dans le(s) dictionnaire(s) commençant par la chaîne *Il a*, on évite d'examiner les chaînes plus longues commençant par ces deux mots). Chaque fois que l'on trouve une entrée dans un dictionnaire identique à l'une de ces chaînes, le système conserve l'unité correspondant à cette entrée : le but est de constituer au moins une liste de telles unités correspondant à une segmentation complète du texte. Si l'entrée est un amalgame, on rajoute plusieurs unités dans la liste. Si plusieurs choix sont possibles, on crée (virtuellement) autant de listes qu'il le faut.

Dans notre exemple, si l'on suppose que l'on utilise un analyseur morphologique sans connaissances syntaxiques, au moins une dizaine de listes différentes seront ainsi produites (aucune d'entre elles n'étant d'ailleurs exactement identique à la liste que nous proposions au début de ce chapitre). En effet *Il* correspondra à une unité u_1 (peut-être deux, si le pronom personnel et le pronom impersonnel sont deux unités distinctes), *a* et *mangé* correspondront sans doute chacun à une unité, resp. u_2 et u_3 (nous avons vu les difficultés posées par les temps composés), *des* à au moins deux possibilités : u_4 (déterminant) et $u_4' + u_4''$ (préposition + déterminant). La suite de la phrase conduira à deux segmentations différentes : une unité u_5 pour le mot composé *pommes de terre* d'une part, et d'autre part une décomposition en *pommes* (au moins deux possibilités u_6 et u_6' : il existe un verbe *pommer* !),

de (une unité u_7), et *terre* (là encore, deux possibilités u_8 et u_8' : n'oublions pas le verbe *terrer*, qui peut d'ailleurs correspondre à lui tout seul à quatre possibilités, si la 1è et la 3è personne de l'indicatif et du subjonctif sont considérées comme des unités distinctes). Cela fait au minimum dix listes possibles... Généralement, toutes ces listes ne sont pas représentées individuellement dans le système : on préfèrera les coder sous forme d'un graphe, tel que celui présenté ci-dessous (l'unité u_9 correspond au point qui termine la phrase).



il a mangé des pommes de terre

La question des flexions peut modifier le principe de base que l'on vient d'exposer. Traditionnellement, on a un dictionnaire de radicaux, qui comporte aussi la conjugaison utilisable, et un dictionnaire de désinences. L'**algorithme** (c'est souvent à celui-ci que l'on se réfère quand on parle d"analyseur morphologique") consiste à découper le mot de toutes les façons possibles pour obtenir toutes les combinaisons acceptables radical + désinence(s) (cf. D. Kayser, 1985 p.1200, montrant comment le mot *régions* pourrait théoriquement faire l'objet de 7 découpages différents).

De nos jours, l'augmentation des mémoires centrales fait préférer de plus en plus l'utilisation directe d'un **dictionnaire de formes fléchies**, généré automatiquement par application des règles de flexions sur le dictionnaire des formes lemmatisées, avec référence à ce dictionnaire, qui comporte toute l'information attachée au mot (à l'exception bien sûr de l'information flexionnelle).

Quoi qu'il en soit de cette question, on peut dire qu'en matière de traitement automatique, les faits de flexion constituent, depuis les années 60, le secteur le mieux maîtrisé de la morphologie.

3.2. Le traitement des ambiguïtés morphologiques

Comme on l'a vu sur l'exemple ci-dessus, les ambiguïtés morphologiques peuvent concerner le découpage et / ou la catégorie du mot (ainsi *figure* est soit un substantif sans désinence soit un verbe suivi d'une désinence), et aussi l'assignation de la valeur de la désinence flexionnelle (ainsi *figure* traité comme verbe conjugué a une désinence *-e* qui est doublement ambiguë : présent de l'indicatif ou du subjonctif, 1° ou 3° personne du singulier).

Ces cas d'ambiguïté, en particulier ceux de **polycatégorie**, conduisent à un nombre considérable d'ambiguïtés artificielles : on a pu facilement s'en convaincre sur l'exemple — tout à fait banal — que nous avons présenté. Ceux-ci ne sont pas décidables dans le cadre d'un analyseur strictement morphologique, car la levée de ce type d'ambiguïté ne peut se faire qu'en prenant en considération le **contexte** du mot. Très souvent, des connaissances syntaxiques locales sur l'environnement immédiat du mot dans la phrase suffisent pour lever l'ambiguïté : ainsi dans *je bois*, le terme *bois* ne peut être qu'un verbe. Dans d'autres cas en revanche, l'environnement à considérer est plus étendu : ainsi dans *le bois* l'incertitude demeure sur la catégorie des deux termes et n'est levée que si l'on a, par exemple, *le bois était touffu* par opposition à *je le bois avec plaisir*. Il arrive parfois que le contexte syntaxique entier de la phrase soit insuffisant pour lever toutes les ambiguïtés catégorielles : dans de tels cas, il y a non seulement ambiguïté morphologique (ambiguïté locale sur un ou plusieurs mots) mais **ambiguïté syntaxique** de la phrase (ambiguïté globale de toute la structure qui peut être analysée de plusieurs façons différentes). Evoquons ici quelques exemples célèbres, qui faisaient déjà les délices des spécialistes de traduction automatique il y a plusieurs décennies : *Le pilote ferme la porte* ou *Le cuisinier sale la note* ("article + substantif + verbe + article + substantif" ou "article + substantif + adjectif + pronom personnel objet + verbe") ; *La petite ferme le voile* ou *La petite brise la glace* ("article + adjectif + substantif + pronom personnel objet + verbe" ou "article + adjectif substantivé + verbe + article + substantif").

Deux grands types de solutions sont possibles :

- ou bien l'on recourt, dès le niveau morphologique, à une analyse syntaxique minimale permettant d'effectuer une désambiguïsation sur des bases statistiques de cooccurrence de catégories : certains analyseurs morphologiques adoptent ainsi la stratégie de choisir parmi les diverses catégories possibles d'un mot celle qui est la plus plausible dans le contexte immédiat, c'est-à-dire en fonction de la catégorie du (ou des) mot(s) précédant immédiatement le mot ambigu (en se fondant par exemple sur une matrice d'échantillonnage probabiliste constituée à partir de l'analyse de divers corpus). Ainsi, l'analyseur morphologique développé par le CRISS (cf. G. Lallich-Boidin *& al.* 1990) utilise l'information catégorielle des deux mots qui précèdent le mot ambigu pour décider de la catégorie à lui attribuer : appelons A et B les catégories de ces deux mots, et C et C' les deux catégories possibles du mot ambigu ; l'analyseur se reporte à une table où sont consignées les probabilités d'occurrences des triplets (A B C) et (A B C') et choisit la catégorie correspondant au triplet de plus forte probabilité. Le principe de base utilisé est connu sous le nom de chaîne de Markov. (En fait, l'algorithme est plus complexe que ne le laisse entendre le cas que nous venons de décrire : il faut en effet tenir compte des cas où plusieurs mots ambigus se succèdent et calculer, dans le graphe des chemins possibles, le chemin de plus forte probabilité globale.)

- ou bien l'on se contente, au niveau morphologique, de noter la pluri-appartenance catégorielle et l'on attend des informations d'un autre niveau pour lever l'ambiguïté : selon les cas, il faudra des informations syntaxiques locales, ou des informations syntaxiques plus étendues, parfois même la polycatégorie ne pourra pas être résolue dans le cadre de la syntaxe de la phrase. Les modalités de recours à des informations d'un autre niveau (ici des informations syntaxiques pour la résolution d'ambiguïtés morphologiques) dépendent de l'**architecture** d'ensemble du système : si l'on a affaire à un système strictement **modulaire**, la sortie de l'analyse morphologique complète (avec ses solutions multiples en cas d'ambiguïté) constitue l'entrée de l'analyse syntaxique (qui doit alors éliminer les solutions parasites du point de vue des règles syntaxiques) ; s'il y a **coopération** des niveaux, l'analyse morphologique peut s'interrompre à chaque ambiguïté détectée et faire prendre ponctuellement le relais à la syntaxe puis reprendre. Cette coopération peut prendre aussi une autre forme : on peut fondre l'analyse morphologique dans l'analyse syntaxique. Dans ce cas, la recherche des unités se limite aux unités dont la catégorie correspond aux attentes de l'analyseur syntaxique. Ainsi, pour analyser *je bois*, un analyseur syntaxique limité aux trois règles de production (cf. *infra* chapitre 4) suivantes :

PH —> GN Verbe

GN —> Pronom

GN —> Déterminant + Nom

guiderait l'analyse morphologique de la façon suivante :

1) recherche du début d'un GN : seules les unités de catégorie Pronom ou Déterminant sont recherchées : *je* est identifié comme pronom, ce qui termine la recherche du GN ;

2) recherche d'un verbe : des deux catégories possibles pour *bois*, seule la forme verbale est recherchée par l'analyseur. L'ambiguïté est donc levée immédiatement.

3.3. Les traitements avec "incertitudes"

Plusieurs types d'objectifs peuvent conduire à modifier plus ou moins radicalement ce qui vient d'être dit. Citons en particulier les **correcteurs orthographiques** et plus généralement les **analyseurs tolérants** quant aux fautes d'orthographe : il se trouvent confrontés à des mots inconnus dans le dictionnaire, à des flexions erronées, à des fautes d'accent, etc., qui rendent encore plus problématique l'autonomie d'un niveau strictement morphologique ; dans ce cas, il faut à tout prix lier morphologie et syntaxe (cf. *infra* chapitre 4).

Citons également les traitements avec **dictionnaire incomplet** (par exemple pour l'interrogation de bases documentaires). L'objectif est alors

d'analyser un énoncé contenant des termes (en particulier des termes techniques) inconnus du système. Le système cherche à trouver la catégorie syntaxique d'un mot inconnu en s'appuyant sur deux types de connaissances : d'une part, des données sur les terminaisons (un mot se terminant par *-ation* aura bien des chances d'être un nom) et d'autre part l'environnement syntaxique du mot ; comme pour les correcteurs orthographiques, les algorithmes d'analyse morphologique doivent donc être intimement liés aux processus d'analyse syntaxique.

Un autre cas de traitement avec connaissances incomplètes concerne les textes saisis en majuscules : l'absence de distinction minuscules / majuscules et surtout l'absence de lettres accentuées augmente considérablement le nombre d'ambiguïtés artificielles ; ces conditions sont proches, là encore, de celles de la correction orthographique.

4. Outils et techniques informatiques pour l'analyse morphologique

Avant d'aborder les techniques informatiques, nous évoquerons tout d'abord l'existence de **dictionnaires électroniques.** On désigne sous ce terme des dictionnaires qui sont élaborés spécialement pour le traitement automatique de la langue. De tels dictionnaires ne doivent pas être confondus avec les **dictionnaires d'usage** présentés sur support électronique (tels par exemple le Robert pour le français, ou le Longman pour l'anglais), qui ne sont que des variantes des versions classiques sur papier. Les véritables dictionnaires électroniques, au contraire, sont conçus selon des principes de systématicité et d'exhaustivité qui en font des outils directement utilisables dans le cadre d'un traitement automatique.

Pour le français, mentionnons l'existence de deux gros dictionnaires électroniques comportant des informations morphologiques : le DELAS élaboré par l'équipe de M. Gross à l'Université Paris VII, et BDLEX élaboré par l'équipe de G. Perennou à l'Université de Toulouse, dans le cadre du GRECO "Communication parlée".

Le **DELAS** est un dictionnaire de "mots simples", le mot étant défini ici de manière rigoureusement graphique : tout signe de ponctuation est exclu (ainsi *aujourd*, *hui*, *jusqu*, *parce* sont des entrées distinctes du DELAS). Typiquement, une entrée du DELAS consiste en un mot sous sa forme lemmatisée, suivi d'un "code morphologique" constitué de sa catégorie grammaticale (l'une des 8 catégories usuelles : nom, adjectif, verbe, adverbe, déterminant, préposition, conjonction, interjection) et d'un numéro renvoyant au modèle flexionnel associé à ce mot. Si l'entrée est polycatégorielle, elle est suivie de plusieurs codes morphologiques. Par ailleurs un code spécial est

attribué aux unités non autonomes comme *aujourd*. A noter aussi que quelques marqueurs spécifiques sont rajoutés au code pour indiquer certaines particularités : par exemple le fait pour un verbe d'être impersonnel ou défectif (un fichier externe précise les flexions manquantes pour chaque verbe défectif). Ainsi on trouvera dans le DELAS les entrées suivantes :

château,.N3
colmater,.V3
frire,.V90D
déjeuner,.N1.V3
avant,.N1.A80.PREP.ADV
parce,.XINC

qui indiquent que *château* est un nom de la classe flexionnelle n°3 (prend un *x* au pluriel), que *colmater* est un verbe de la classe de conjugaison n°3 (modèle : *chanter*), que *frire* est un verbe de la classe de conjugaison n°90, mais qui est défectif, que *déjeuner* est soit un nom soit un verbe, que *avant* peut être nom, adjectif (invariable : classe n°80), préposition ou adverbe, et enfin que *parce* n'est pas une unité autonome.

Le DELAS contenait en 1990 près de 80 000 entrées. Ce n'est qu'un élément parmi tout un système de dictionnaires comprenant, entre autres :

• le DELAF, dictionnaire des formes fléchies associées aux entrées du DELAS : ce dictionnaire est généré automatiquement à partir du DELAS. Il comportait en 1990 plus de 600 000 entrées, qui se présentent de la manière suivante :

maisons,maison.N21: Nfp
irais,aller.V16 : CPr1s : CPr2s

qui indiquent que *maisons* est une forme fléchie de *maison* (code N21), qui marque le féminin pluriel, et que *irais* est une forme fléchie de *aller* (code V16), qui marque soit la 1ère personne soit la 2ème personne du singulier du conditionnel présent.

• le DELAC, dictionnaire de mots composés, comportant pour l'instant des mots composés insécables, essentiellement des adverbes, des conjonctions de subordination et des noms (de forme A N, N A, N *à* N, N *de* N, N N, P N ou encore V N).

• le DELAP, dictionnaire "phonémique", qui associe à chaque entrée du DELAS une représentation en phonèmes à partir de laquelle on peut générer, à l'aide d'algorithmes simples, une transcription phonétique du mot. Ce dictionnaire permet aussi de construire le DELAP-F, qui contient la représentation phonémique des formes fléchies correspondantes.

BDLEX est un dictionnaire électronique plus orienté vers le traitement de l'oral (cf. *supra,* chapitre 1). Dans ce but, à chaque entrée sont associées une série d'informations précises : en plus de la représentation phonologique

syllabée, on trouve des indications sur le fonctionnement phonologique de la finale, sur le nombre d'homophones du mot, une représentation phonologique et syllabique en classes majeures, etc. Les catégories grammaticales sont à peu près les mêmes que dans le DELAS, à part l'introduction de classes mixtes Nom-Adjectif. Les mots composés, en revanche, sont des entrées de BDLEX au même titre que les mots simples, si ce n'est que leurs éventuelles flexions sont répertoriées dans un fichier séparé. Autre différence, les mots polycatégoriels comme *déjeuner* font l'objet de deux entrées distinctes. La couverture de ce dictionnaire était sensiblement moins grande que celle du DELAS en 1990.

En matière de traitement automatique de faits dérivationnels, signalons l'élaboration (en cours) d'un "**dictionnaire dérivationnel** du français" à l'Université de Lille III, sous la direction de D. Corbin et P. Corbin (cf. leur article de 1991).

Enfin de gros projets sont actuellement en cours pour constituer de véritables **bases de données lexicales** comportant aussi bien des informations sémantiques que des informations morphologiques et syntaxiques. Ainsi le projet européen GENELEX (un des plus gros projets EUREKA de ces dernières années) s'est donné pour objectif de constituer de telles bases pour le français, l'espanol et l'italien, et il réunit pour cela des éditeurs, des entreprises informatiques et des laboratoires universitaires (pour la France : Hachette, Bull, GSI-ERLI, SEMA-GROUP, et le LADL) ; cf. M. Nossin 1991.

Du point de vue des **techniques informatiques de reconnaissance des formes fléchies**, la situation a beaucoup évolué. Les premiers analyseurs morphologiques (comme par exemple celui du programme SHRDLU de Winograd 1972) utilisaient une analyse procédurale, fondée sur la notion d'"automate à nombre fini d'états" : l'analyseur "lisait" les lettres du mot une à une en partant de la dernière, et changeait d'état en fonction de la lettre lue. A certains de ces états était associée une terminaison valide du dictionnaire des flexions, et l'analyseur regardait alors dans un dictionnaire de radicaux si un radical de la classe flexionnelle adéquate existait ; si c'était le cas, une solution était trouvée. Pour prendre un exemple simple, soit à analyser le mot *chevaux* : le programme se trouve au départ dans l'état 0. La lecture du *x* le place dans un nouvel état, par exemple l'état 5, auquel est associée une terminaison de pluriel existante : le logiciel va donc chercher si le radical *chevau* existe. Il échoue, et lit donc la lettre suivante *u*. Il se retrouve du même coup, dans un nouvel état, mettons 14, auquel est associée une terminaison valide. Il cherche dans le dictionnaire le radical *cheva*, le trouve, et peut identifier *chevaux* comme étant le pluriel de *cheval*. Une solution est donc obtenue. Le processus se poursuit néanmoins à la recherche d'autres solutions éventuelles, mais aucune terminaison valide n'étant associée aux états suivants de l'automate, aucune autre solution n'est trouvée. On trouvera une description plus détaillée de ce type d'algorithme dans G. Sabah (1989, pp. 25-28).

Ce type d'approche est dit **"procédural"**, parce qu'une partie des connaissances linguistiques (celle qui concerne les terminaisons) est directement codée dans l'algorithme lui-même (sous forme d'états différents de l'analyseur). Cette technique a été rapidement remplacée par une approche dite **"déclarative"**, dans laquelle toutes les connaissances linguistiques sont séparées de l'algorithme lui-même : les radicaux, les classes de flexions, et les listes de terminaisons correspondantes sont décrites dans des structures de données autonomes, et l'algorithme consiste à découper le mot de toutes les manières possibles en comparant chaque fois les deux parties du mot aux données. Cette méthode a l'avantage de séparer le travail de description linguistique de l'utilisation de ces données : le même algorithme peut donc être mis en oeuvre sur plusieurs jeux de données (des langues différentes par exemple), et l'acquisition et l'enrichissement des données (notamment la correction des erreurs) sont beaucoup plus aisées ; en particulier, cela permet à un linguiste d'entrer ces données sans avoir à maîtriser l'ensemble du programme. On trouvera là aussi une description plus détaillée de cette technique dans G. Sabah (1989, pp. 28-35).

De nos jours, l'augmentation des mémoires d'ordinateur fait préférer généralement une méthode encore plus simple : on génère une fois pour toutes le **dictionnaire des formes fléchies**, et l'algorithme consiste simplement à aller chercher dans ce dictionnaire la ou les forme(s) correspondant au mot analysé. Ce n'est que dans le cas où le mot n'est pas trouvé que l'on va chercher, par des méthodes qui s'appuient à la fois sur le type de terminaison et sur l'environnement syntaxique, à suppléer cette absence d'information dans le dictionnaire pour essayer de deviner au moins la catégorie syntaxique du mot (voir *supra* § 3.3. : connaissances incomplètes).

On trouve cependant d'autres méthodes encore à l'oeuvre : ainsi l'analyseur du CRISS utilise une méthode procédurale assez originale pour traiter des flexions (cf. G. Lallich-Boidin *& al.* 1990). D'une part, il opère une régularisation systématique des flexions (ainsi le mot *chevaux* sera-t-il systématiquement remplacé par *chevaus*, puis par *chevals*, la seule désinence régulière du pluriel étant un *s* final), avant de rechercher le radical correspondant. D'autre part, la recherche des flexions utilise l'ordre dans lequel elles apparaissent en français : ainsi puisque les flexions de genre précèdent celles de nombre dans les adjectifs, l'analyseur tire parti de cette règle pour chercher d'abord à déterminer le nombre, puis le genre (puisqu'il opère une lecture des lettres de droite à gauche). De même pour les verbes, le nombre, la personne, le temps seront analysés dans cet ordre (*chanterions* est ainsi effectivement décomposé en *chant-er-i-ons*).

5. Perspectives

Comme on aura pu s'en convaincre après ce tour d'horizon, les problèmes d'ordre morphologique sont loin d'être entièrement résolus dans le domaine du traitement automatique. Si la question des flexions grammaticales est, depuis longtemps, complètement maîtrisée, il reste toute une série de voies de recherche encore largement ouvertes : qu'il s'agisse de la prise en compte des phénomènes de dérivation et de composition lexicales, ou de l'analyse avec connaissances incomplètes (soit pour traiter les mots nouveaux, qui apparaissent tous les jours dans la langue, soit pour réaliser des correcteurs d'orthographe). Ce sont autant de défis que linguistes et informaticiens doivent relever par un travail en commun, par l'approfondissement des connaissances théoriques sur le sujet et par la mise en oeuvre de techniques informatiques nouvelles appropriées à chacun de ces problèmes.

Catherine FUCHS et Bernard VICTORRI
(ELSAP-CNRS)

Repères bibliographiques

1. Présentations d'ensemble :

KAYSER, D. (1985) : Des machines qui comprennent notre langue, *La Recherche*, 16 :170, 1198-1212.

[Article de vulgarisation consacré aux problèmes de la compréhension automatique ; sur l'analyse morphologique, voir p. 1200.]

SABAH, G. (1989) : *L'intelligence artificielle et le langage* (vol. 2 : "Processus de compréhension"), Paris, Hermès.

[Voir le ch. 1 : "L'analyse morphologique".]

SMITH, G. (ed.) (1991) : *Computers and human language*, San Mateo, Kaufmann.

[Voir le ch. 1 : "Components of words".]

2. Descriptions et théories linguistiques :

BESCHERELLE (réédition 1990) : *L'art de conjuguer*, Paris, Hatier.

[Le "classique" des flexions de conjugaison du verbe français : couvre environ 7.000 entrées.]

CORBIN, D. (1987) : *Morphologie dérivationnelle et structure du lexique*, Tübingen, Niemeyer, 2 vol.

[Réflexion théorique linguistique sur les faits de dérivation en français, et proposition d'un "modèle lexical stratifié", avec exemples d'applications : suffixations adverbiales en *-ment*, suffixations nominales en *-ette*, préfixations en *dé-* et en *anti-*.]

CORBIN, D. (ed) (1991) : "La formation des mots : structures et interprétations", *Lexique*, 10, Lille, Presses Universitaires.

[Série d'études préparatoires à l'élaboration d'une grammaire et d'un dictionnaire dérivationnel du français.]

DUGAS, A. & MOLINIER, Ch. (eds) (1992) : "La productivité lexicale", *Langue Française*, 96, Paris, Larousse.

GROSS, G. (1990) : Les mots composés, *Modèles Linguistiques*, XII :1, Lille, 47-63.

[Etude des degrés de figement des mots composés, dans la perspective de l'équipe de M. Gross.]

JACQUEMIN, Ch. & CADIOT, P. (1992) : Approche connexionniste de la composition nominale, *Actes du Congrès Neuro-Nîmes 92*, E.C.2, 333-345.

[Exemple d'un traitement connexionniste de la composition nominale en français, appliqué aux noms composés en "N *à* N" et aux groupes en "Prép. Nloc.".]

MARTINET, A. (1960, rééd. 1991) : *Eléments de linguistique générale*, Paris, Colin.

[Ouvrage de référence en matière d'analyse structuraliste de la langue, et de découpage en termes de morphèmes.]

3. Dictionnaires électroniques :

CORBIN, D. & CORBIN, P. (1991) : Vers le Dictionnaire Dérivationnel du Français, *Lexique*, 10, Lille, Presses Universitaires, 147-161.

[Présentation de ce dictionnaire en cours d'élaboration, illustré sur l'exemple de l'entrée *"poivrier"*.]

COURTOIS, BL. & SILBERZTEIN, M. (eds.) (1990) : "Dictionnaires électroniques du français", *Langue Française*, 87, Paris, Larousse.

[Présentation des différents dictionnaires électroniques élaborés dans l'équipe de M. Gross : DELAS, DELAF, DELAC, DELAP.

Sur les mots composés, voir en particulier les articles de M. Silberztein, G. Gross, R. Jung et R. Vivès.]

NOSSIN, M. (1991) : Le projet GENELEX : EUREKA pour les dictionnaires génériques, *Actes du Colloque Génie Linguistique 91*, Versailles, EC2.

PERENNOU, G. *& al.* (1987) : BDLEX, base de données lexicale du français écrit et parlé (rapport du GRECO "Communication parlée").

4. Analyseurs morphologiques automatiques :

HERZOG, O. & ROLLINGER, C.R. (eds.) (1991) : *Text understanding in LILOG*, Berlin, Springer, Coll. "Lecture notes on artificial intelligence".

[Voir pp. 105-182, pour une présentation d'un lexique et d'un analyseur morphologique de l'allemand, avec un traitement particulièrement soigné des mots inconnus et des mots composés (phénomène massif en allemand).]

LALLICH-BOIDIN, G. *& al.* (1990) : Une interface multimode pour une interaction homme-machine avec une base de connaissances ; analyse du français : achèvement et implantation de l'analyseur morpho-syntaxique, Rapport technique ESPRIT, Grenoble.

[Présentation de l'analyseur morpho-syntaxique du CRISS dans ses aspects linguistiques, formels et algorithmiques.]

WINOGRAD, T. (1972) : *Understanding natural language*, San Diego, Academic Press et Edinburgh University Press.

[Voir le § 3.9., pp.73-76, pour une présentation d'un programme "historique" d'analyse procédurale des terminaisons flexionnelles des mots en anglais.]

4
SYNTAXE

En traitement automatique, l'analyse syntaxique consiste à associer à la chaîne découpée en unités, une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités.

Cet objectif ne constitue pas un but en soi. Pour certains traitements, l'analyse syntaxique est un **outil** permettant la mise en oeuvre d'une application particulière (par exemple pour l'élaboration de correcteurs de fautes d'orthographe dues à des faits de syntaxe comme les phénomènes d'accord) ; pour d'autres elle peut constituer un **préalable** à l'analyse sémantique (par exemple en vue de la traduction ou de la compréhension automatique de textes) : elle intervient alors comme un **intermédiaire** entre l'analyse morphologique et l'analyse sémantique.

Il convient toutefois d'insister sur le fait que certains traitements automatiques récusent, de fait, la **nécessité** d'une étape syntaxique. Il existe par exemple des systèmes de compréhension automatique qui font l'économie d'une telle étape (comme certains systèmes assez élémentaires d'interrogation de bases de données reposant uniquement sur le repérage de "mots clés"). En contrepartie, de tels systèmes sont contraints de recourir à une sémantique entièrement dépendante du domaine d'application. Seul en effet le passage par une étape syntaxique peut permettre de construire des systèmes suffisamment généraux pour ne pas être totalement dépendants d'une application particulière.

Précisons par ailleurs que les systèmes qui recourent à une analyse syntaxique n'en font pas nécessairement pour autant un **module** de traitement indépendant : certains ne posent pas de coupure nette entre analyse morpho-

logique et analyse syntaxique (cf. *infra*, § 4.1.), d'autres subordonnent l'analyse syntaxique à l'analyse sémantique (cf. *infra*, chapitre 5, § 3.1.).

Dans la plupart des théories syntaxiques, la structure syntaxique est conçue de façon **hiérarchique** : les unités sont regroupées en constituants intermédiaires de taille variable (syntagmes), qui s'emboîtent les uns dans les autres ; c'est ce qui explique que la représentation en sortie de l'analyseur soit généralement un **arbre**, l'information syntaxique étant attachée aux noeuds et aux branches de cet arbre.

Reprenons l'exemple de la chaîne :

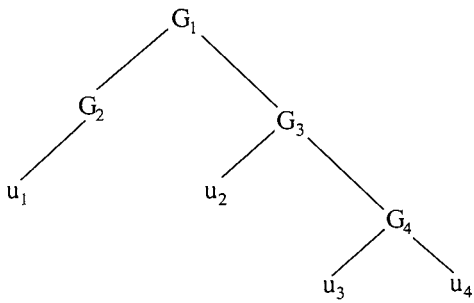
Jean a mangé des pommes de terre.

et sa représentation morphologique

(u_1,u_2,u_3,u_4,u_5)

obtenue au chapitre précédent.

La sortie de l'analyseur syntaxique pourra être par exemple l'arbre suivant (u_5, qui a servi à délimiter la phrase, n'y figure pas) :



avec (entre autres) les informations suivantes :

G_1 : unités : (u_1,u_2,u_3,u_4) (*Pierre a mangé des pommes de terre*)
catégorie : phrase
constituants : G_2 : sujet, G_3 : tête

G_2 : unités : (u_1) (*Pierre*)
catégorie : groupe nominal
constituants : u_1 : tête

G_3 : unités : (u_2,u_3,u_4) (*a mangé des pommes de terre*)
catégorie : groupe verbal
constituants : u_2 : tête, G_4 : complément

G_4 : unités : (u_3,u_4) (*des pommes de terre*)
catégorie : groupe nominal
constituants : u_3 : spécificateur, u_4 : tête.

Si l'analyse morphologique est aujourd'hui un problème relativement bien résolu en traitement automatique (du moins en ce qui concerne les flexions), en revanche ce n'est pas le cas de l'analyse syntaxique, malgré des efforts considérables. Bien des progrès ont été effectués ces dernières années, mais il n'existe pas encore d'analyseur **général** de la langue (en tout cas pour le français), c'est-à-dire dont la couverture soit suffisamment large pour que l'on puisse être sûr que toutes les tournures de la langue soient traitables avec les moyens proposés : aucun analyseur syntaxique ne semble capable, à l'heure actuelle, d'analyser un texte français tout venant sélectionné au hasard (passage de roman, article de journal,...). La plupart des systèmes font des choix sur les types de phénomènes syntaxiques traités (sans parler des limitations dues au lexique), choix qui sont fonction à la fois des objectifs du système et de ce que l'on sait théoriser d'une façon implémentable. Signalons toutefois l'existence d'analyseurs qui traitent effectivement des textes tout venant, mais de manière "approximative" : c'est le cas par exemple de l'analyseur de T. Strzalkowski (1992) pour l'anglais, qui a traité des corpus imposants (en tout, environ 50 millions de mots, à la vitesse moyenne de plus de 2000 mots par minute !), mais en produisant une proportion importante d'analyses incorrectes.

Selon les **objectifs** du système, les types de structures syntaxiques et la nature des données que l'on traite varient. En ce qui concerne les structures syntaxiques, rappelons par exemple que certains analyseurs restreignent *a priori* les types de phrases qu'ils traitent (c'est le cas notamment des analyseurs syntaxiques limités contenus dans certains systèmes d'interrogation de bases de données, qui imposent à l'utilisateur le recours à des schémas de phrases simples, généralement à tête nominale). En ce qui concerne la nature des données, les systèmes acceptent de traiter des phrases plus ou moins grammaticales, selon leurs objectifs. A cet égard, il importe de distinguer deux sortes de **tolérance** : d'une part on peut avoir besoin d'opter pour une grammaire plus ou moins rigide, pour une analyse plus ou moins tolérante à l'égard de phrases non totalement grammaticales, mais attestées et acceptables pour des sujets humains (il est clair par exemple que pour un système de dialogue homme-machine, il est capital de pouvoir ne pas rejeter des requêtes formulées de façon approximative, non canonique, etc. : cf. *infra*, chapitre 10) ; c'est ce niveau de tolérance, difficile à obtenir pour un analyseur, que s'efforcent d'atteindre les systèmes les plus évolués. D'autre part, on peut se donner la facilité, du moins pour un analyseur destiné à la reconnaissance, d'accepter des phrases a-grammaticales, dont on sait qu'on ne les rencontrera jamais, comme par exemple * *Je me me donne* (les exigences en matière de grammaticalité seront en revanche beaucoup plus contraignantes s'il s'agit d'une syntaxe destinée à la génération : cf. *infra*, chapitre 9).

Par ailleurs, au plan de la théorisation implémentable, l'absence actuelle d'analyseur syntaxique véritablement général s'explique par deux ordres de facteurs :

• d'abord et avant tout, par la complexité des phénomènes syntaxiques : si ceux-ci font l'objet de descriptions fines (bien que souvent parcellaires) de la part des linguistes, en revanche les théories formelles sont encore loin d'en rendre compte dans toute leur complexité ;

• mais aussi par la relative inadéquation entre les théories syntaxiques (qui recourent aux représentations structurelles pour décrire et expliquer les régularités observées) et la tâche spécifique de reconnaissance automatique (où l'on doit construire une représentation à partir de la seule donnée de la chaîne linéaire).

Cette situation nous conduit à présenter en quatre étapes la problématique de l'analyse syntaxique en traitement automatique : nous aborderons tout d'abord les principaux phénomènes syntaxiques dont doivent rendre compte les analyseurs, en particulier ceux dont la représentation pose problème pour toute théorie syntaxique formalisée (§ 1.) ; nous considèrerons ensuite les difficultés inhérentes à la tâche d'analyse automatique, c'est-à-dire au passage de l'information d'entrée à la représentation de sortie (§ 2.) ; puis nous évoquerons les divers formalismes linguistiques utilisés aujourd'hui en traitement automatique de la syntaxe (§ 3.) ; enfin nous présenterons sous divers aspects les principes mis en oeuvre dans les analyseurs (§ 4.).

1. Les phénomènes syntaxiques

Nous ne ferons ici que rappeler certains principes de base.

1.1. La phrase

En pratique, la plupart des analyseurs syntaxiques sont des analyseurs de **phrases**. Les raisons de ce privilège accordé à la phrase sont diverses. D'une part on sait mal décrire (et, *a fortiori*, traiter automatiquement) des unités de taille supérieure. D'autre part, à la suite d'une longue tradition linguistique, on s'accorde à reconnaître à la phrase une spécificité de fonctionnement qui ne se trouve pas dans les unités de taille inférieure (ni dans le mot ni dans le syntagme) : c'est en effet au niveau de la phrase que se construit la prédication (mise en relation d'un sujet et d'un prédicat) et, corrélativement, l'assertion. Bien entendu, le choix de construire des analyseurs de phrases oblige, pour le traitement automatique des textes, à aborder ultérieurement la question des liens entre les phrases (cf. *infra*, chapitre 8).

Au plan **théorique**, la caractérisation de la phrase rencontre un certain nombre de problèmes. Il est classique de considérer que la phrase est constituée d'une ou plusieurs **propositions** et que la proposition est elle-même repérable à la présence d'un **verbe** conjugué. Toutefois, à côté des cas indis-

cutables, où l'articulation des propositions au sein de la phrase est traduite par des marqueurs spécifiques et où chaque proposition comporte un verbe (proposition complétive enchâssée dans la principale : *Il est ravi qu'on ait fait appel à lui*, proposition circonstancielle subordonnée à la principale : *Il est ravi parce qu'on a fait appel à lui*, propositions coordonnées : *Il est ravi car on a fait appel à lui*), il existe des cas-limites, du fait de la nature moins nette de certains marqueurs (ainsi dans : *Il est ravi : on a fait appel à lui,* ou dans : *Il est ravi, bien sûr, on a fait appel à lui,* les deux propositions constituent-elles deux phrases distinctes ou une seule ?), ou de l'absence de verbe (ex : *Allo!* ; *Pourquoi donc ?* ; *Lui, un assassin? Impossible !*). La phrase, on le sait, ne coïncide pas toujours avec l'**énoncé** (unité minimale d'énonciation).

Si l'on se tourne à présent vers des critères de nature formelle, il apparaît que l'identification de la phrase sur des bases purement **typographiques** se heurte, elle aussi, à un certain nombre de difficultés. A l'écrit, une caractérisation de la phrase comme suite de mots comprise entre un point initial (avec une majuscule sur l'initiale de la première lettre du premier mot) et un point final (ou entre deux signes de ponctuation forte, comme le point d'interrogation ou le point d'exclamation) s'avère en effet n'être une condition ni nécessaire ni suffisante.

Elle n'est pas nécessaire, dans la mesure où il est possible de traiter comme des phrases distinctes, à part entière, les séquences séparées par les signes deux points, guillemets ou parenthèses dans des exemples comme : *Cette machine est lente : c'est un vieux modèle* ; *La notice indique : "Cet appareil ne doit pas rester branché en permanence"* ; *Cliquez sur les bords du guide (les guides sont représentés par des lignes en pointillés), sur la position désirée.*

Elle n'est pas suffisante non plus, car le point, le point d'exclamation et le point d'interrogation apparaissent à l'intérieur de séquences que l'on peut être tenté de traiter comme des phrases uniques : c'est évidemment le cas lorque le point se trouve après une abréviation (ex : *M.* pour *Monsieur* dans une phrase comme *Cette lettre est adressée à M. Dupont*) ; de façon moins indiscutable, citons les cas où le point d'exclamation ou d'interrogation apparaît après certains mots correspondant à des unités de communication (ex : *Une erreur, peut-être ?, qui ressemble à une malveillance* ; *C'est hélas! irrécupérable* ; *Ouf! je respire*). Rappelons par ailleurs que la majuscule, que supporte la première lettre du premier mot de la phrase, ne constitue pas non plus un indice sûr, puisqu'elle se trouve également à l'initiale de tout nom propre, à l'intérieur de la phrase.

La segmentation d'un texte en phrases est, on le voit, un problème difficile qui, tout comme la segmentation en mots, engage nécessairement un certain nombre de décisions pour lequelles les traitements automatiques se trouvent pris entre les deux impératifs contradictoires déjà signalés : le souci de cohérence théorique d'un côté, et la recherche de solutions opérationnelles au moindre coût de l'autre.

En fait, peu d'attention a été portée à ces difficultés jusqu'à présent. Bien des systèmes les évacuent en supposant un "pré-traitement" manuel du texte (reposant sur des choix plus ou moins explicites), qui permet d'identifier les phrases sur des bases purement formelles typographiques. D'autres utilisent des algorithmes simples, qui, sans résoudre complètement le problème, se montrent raisonnablement efficaces. Ainsi, dans le projet LILOG, développé par IBM-Allemagne (cf. *infra*, chapitre 8), l'analyseur essaie de traiter toute chaîne de caractères comprise entre deux séparateurs potentiels de phrase (y compris donc les points qui marquent des abréviations ou des nombres ordinaux en allemand) ; si pour l'une de ces chaînes l'analyse échoue, le système complète la chaîne en allant jusqu'à la marque suivante dans le texte, essaie d'analyser la chaîne ainsi obtenue, et ainsi de suite (cf. O. Herzog & C.R. Rollinger (eds.), 1991, p. 81).

1.2. Syntaxe de la phrase : brefs rappels linguistiques

L'analyse syntaxique de la phrase suppose l'identification des syntagmes, et la structuration hiérarchique de ces syntagmes (en liaison avec la caractérisation de leurs fonctions).

Le **syntagme** est un groupe d'unités, dominé par une "tête" dont la catégorie donne son nom au syntagme (syntagme nominal, syntagme verbal, syntagme adjectival, etc.), qui occupe une certaine position sur la chaîne, et joue un rôle fonctionnel donné. Le syntagme est dit **minimal** lorsqu'il n'est constitué que des catégories obligatoires qui le définissent (ainsi un déterminant + un nom pour le syntagme nominal, ex : *le cordon*) ; il peut au contraire comporter des **ajouts** facultatifs (ainsi un adjectif, un groupe prépositionnel ou une relative, qualifiant le nom ; ex : *le cordon gris ; le cordon le plus ancien ; le cordon de l'imprimante ; le cordon qui relie l'ordinateur à l'imprimante*).

Un syntagme peut donc contenir d'autres syntagmes, qui eux-mêmes en contiennent d'autres, et ainsi de suite : on a affaire à une **structure hiérarchique** qui est à l'origine des représentations de la phrase par un **arbre syntaxique** (cf. l'exemple présenté au début de ce chapitre), dont nous verrons un peu plus loin les insuffisances.

Dans cette optique, la phrase elle-même (tout au moins en français) est composée au moins d'un sujet (généralement un syntagme nominal) et d'un syntagme verbal (le verbe et ses compléments dits "essentiels"). En plus de ces deux éléments, qui constituent la **phrase noyau,** la phrase peut comporter des **ajouts**, notamment des circonstanciels, soit sous forme de syntagmes prépositionnels ou adverbiaux (ex : *Le linguiste écrit la grammaire pendant la journée* ; *Le linguiste écrit la grammaire rapidement*), soit sous forme de propositions — la phrase est alors dite complexe — (ex : *Le linguiste écrit la grammaire quand il en a le temps*) ; rappelons au passage que la distinction

entre "compléments essentiels" du verbe intégrés au groupe verbal, et "compléments circonstanciels" est parfois difficile à opérer (ex : *Il conduit sa voiture de Paris à Marseille / Il vient de Paris / Il va à Marseille ; Il a acheté une voiture à ce marchand / Il a acheté une voiture chez ce marchand*).

Rappelons aussi que l'ordre et la catégorie des constituants occupant une fonction donnée ne sont pas univoques : ainsi, à côté de la phrase "canonique" constituée d'un groupe nominal sujet suivi d'un groupe verbal (ex : *Le linguiste écrit la grammaire*, avec *le linguiste* = GN sujet, et *écrit la grammaire* = GV, lui-même composé de V + GN objet), on trouve, entre autres :

- des phrases où chacune des deux fonctions sujet et objet est remplie (si la construction du verbe le permet) non plus par un GN mais par une proposition P, dite **enchâssée** ; ainsi : *Le linguiste écrit que la grammaire comporte 200 règles ; Que la grammaire comporte 200 règles réjouit le linguiste ; Que la grammaire comporte 200 règles indique que le programme sera long*).

- des phrases dans lesquelles **l'ordre** des constituants est modifié : ainsi dans les cas d'inversion du sujet (*La nuit apparaissent d'étranges créatures*) très fréquents notamment dans les relatives : *L'analyseur qu'a construit cette société est robuste; Le local où est installée cette machine est insalubre.*

1.3. Limites de la représentation hiérarchique

Si la plupart des théories syntaxiques s'accordent *grosso modo* (à la terminologie près) sur les bases que nous venons d'esquisser, elles divergent en revanche sur le traitement d'un certain nombre de phénomènes syntaxiques très fréquents, qui ont en commun d'entrer plus difficilement dans ce cadre hiérarchique de base. A vrai dire, certains d'entre eux rendent même problématique la représentation de la structure syntaxique par un arbre, ce qui n'est pourtant pas remis en question par les théories classiques. En voici une liste (non exhaustive) :

- la **coordination :** qu'il s'agisse de coordination de syntagmes (ex : *Le linguiste et l'informaticien écrivent la grammaire*), ou de propositions (ex : *Le linguiste écrit la grammaire et l'informaticien écrit le programme*), cette relation n'est, par définition, pas hiérarchique puisque les éléments coordonnés sont mis "sur le même plan" (c'est très net pour des coordinations verbales du type *L'informaticien écrit et teste le programme*). Que l'on choisisse de considérer le deuxième élément comme subordonné au premier, ou de créer des catégories spéciales de syntagmes "à plusieurs têtes", ou encore de traiter la coordination comme un mécanisme transversal spécifique, il faut de toutes façons complexifier sérieusement le schéma initial, d'autant plus que la coordination est intimement liée au phénomène de l'ellipse (cf. ci-dessous).

- l'**ellipse**, définie comme "l'effacement" de constituants *a priori* obligatoires (ex : *Le linguiste lit la grammaire, l'informaticien aussi* ;

Quoiqu'aimable, il semblait fatigué ; *Il n'est pas d'accord, moi si*) pose un problème redoutable : faut-il "restituer" les éléments manquants, ce qui peut s'avérer assez délicat et comporter une part d'arbitraire (ainsi faut-il restituer : *Quoiqu'[il fût] aimable* ou *Quoiqu'[il semblât] aimable* dans notre exemple ? — sans parler des ambiguïtés du type : *Paul aime sa femme, et moi aussi* où l'on a trois possibilités : *et moi aussi, j'aime ma femme* ; *et moi aussi, j'aime sa femme* ; *et moi aussi, Paul m'aime*), ou au contraire considérer ces constructions comme tout aussi valides que les constructions canoniques, ce qui complexifie considérablement les règles de syntaxe ? De plus où s'arrête ce phénomène ? Doit-on traiter les exemples suivants comme des ellipses : *Il est plus fort que moi [je ne le suis]* ; *Il s'est conduit comme un imbécile [se conduit]* ?

- les phénomènes d'adjonction de constituants dans des cas comme le **clivage** (ex : *C'est le linguiste qui écrit la grammaire*), l'**apposition** (ex : *Jean, mon voisin, est revenu*), l'**apostrophe** (ex : *Jean, ta femme t'appelle*), les **thématisations** (ex : *Le linguiste, lui, sa grammaire, il l'écrit*), certaines tournures impersonnelles (*Il manque de l'argent* ; *Il pleut des cordes*), etc. Dans chacun de ces cas, la structure hiérarchique que l'on doit choisir n'est pas évidente et la cohérence théorique de ces choix doit être soigneusement pesée en tenant compte, toujours, des impératifs d'efficacité imposés par l'analyse automatique.

- les problèmes de **double portée** ne peuvent être décrits qu'imparfaitement par les structures hiérarchiques. Par exemple, dans la phrase *Il réclame encore des nouilles*, l'adverbe *encore* porte à la fois sur *réclamer* (*réclamer encore* = *réclamer une fois de plus*) et sur *des nouilles* (*encore des nouilles* = *des nouilles supplémentaires*). De même, dans *Je l'ai entendu chanter ce matin*, le circonstanciel de temps porte à la fois sur *entendre* et sur *chanter*. Toute représentation par un arbre, en privilégiant indûment l'attachement à l'un des termes, donne une image faussée de la double relation.

- les **verbes modaux** sont un exemple, entre autres, de difficultés de choix de hiérarchisation. Doit-on, dans une phrase comme *Il peut partir pour Rome*, considérer *pouvoir* comme un modifieur dans un syntagme verbal ayant pour tête *partir*, ou au contraire considérer que le groupe infinitif *partir pour Rome* est complément du verbe *pouvoir* ? La question est d'autant plus difficile à trancher que l'on a affaire, au delà des "vrais" auxiliaires (*Il était / sera parti pour Rome*), à un passage graduel des opérateurs aspectuo-temporels (*Il va / vient de partir pour Rome*), aux modaux proprement dits (*pouvoir*, *devoir*, *vouloir*), puis à des locutions verbales diverses et variées (*avoir l'intention de, rêver de,...*).

- Le continuum de la **lexicalisation**, que nous avons déjà évoqué au chapitre précédent, se prête mal aux choix en tout ou rien qu'impliquent les constructions syntaxiques classiques. En effet, pour toutes sortes d'expressions plus ou moins figées (verbales, nominales, prépositionnelles, ...), il

faudrait pouvoir rendre compte à la fois de la construction interne fine de l'expression, et de l'unité fonctionnelle de la locution qu'elle constitue. Ainsi *au moment où* ; *au moment précis où* devraient pouvoir être analysés en même temps comme des groupes prépositionnels (avec expansion conjonctive), et comme des conjonctions; de même un syntagme comme *un certain nombre d'arguments* mérite une double analyse : l'une dans laquelle *un certain nombre de* est figé comme un simple déterminant de *arguments*, et l'autre dans laquelle *d'arguments* est un groupe prépositionnel complément de *nombre* (comme dans *un nombre important / impressionnant d'arguments*).

2. Les spécificités de la tâche de reconnaissance automatique

Les problèmes que nous venons de décrire doivent être résolus par toute théorie syntaxique, quel que soit son objectif. Mais la tâche spécifique de reconnaissance automatique soulève des difficultés supplémentaires, qui ne sont pas forcément des problèmes importants pour les théories syntaxiques. En effet, pour l'analyse automatique, on ne dispose en entrée que de la **catégorie** et la **position** des unités morphologiques qui ont pu être identifiées, et l'on ne peut pas faire appel à des **tests** (manipulations distributionnelles) impliquant des jugements que seuls les sujets parlants sont en mesure de faire : d'où la nécessité, pour pallier ce manque, d'utiliser des connaissances de toute nature, depuis les règles d'accord jusqu'à des informations sémantiques et des connaissances d'univers. Prenons un exemple pour bien nous faire comprendre. Soient les deux phrases suivantes :

Il a acheté un costume à carreaux

Il a acheté un costume à crédit.

Pour le linguiste, la différence de structure entre ces deux phrases ne pose pas de problème théorique : il peut arguer, entre autres, de leur comportement divergent par rapport à des transformations classiques (le fait que l'on puisse dire *C'est à crédit qu'il a acheté un costume*, mais pas * *C'est à carreaux qu'il a acheté un costume*, et inversement *Un costume à carreaux, il en a acheté un*, mais guère ? *Un costume à crédit, il en a acheté un*) pour conforter son analyse. Ce n'est bien sûr pas le cas pour la machine qui devra posséder une quantité importante d'informations lexicales sur *costume*, *carreau*, *crédit* et *acheter*, pour pouvoir attribuer à chacune des deux phrases la construction correcte.

Ce sont ces problèmes spécifiques que nous allons maintenant passer en revue.

2.1. Problèmes morpho-syntaxiques

Nous avons vu au chapitre précédent les limites d'une analyse morphologique autonome, en particulier pour la désambiguïsation d'unités polycatégorielles, et pour le repérage d'unités discontinues. Ces deux tâches reviennent donc dans la pratique à l'analyseur syntaxique. Cela joue un rôle capital dans la conception même de la stratégie d'analyse, puisqu'elle doit explorer les diverses possibilités laissées ouvertes par le traitement préalable. Cela est particulièrement délicat pour les unités discontinues : qu'il s'agisse de la négation, des temps composés, des comparatifs, etc., il faut repérer des éléments parfois très éloignés sur la chaîne linéaire, et dont la mise en évidence est essentielle à la poursuite de l'analyse. Un exemple suffira à montrer que des règles *ad hoc* plus ou moins locales ne sauraient éviter une approche systématique et globale ; quand on compare les deux phrases :

Depuis qu'il a discuté avec Marie, il est plus convaincu que tu ne le seras jamais

Depuis qu'il a discuté avec Marie, il est plus convaincu que tu ne le feras jamais,

on constate que l'identification de la locution comparative *plus ... que* dans la première phrase ne saurait être préalable à l'analyse de la structure complète de la phrase.

2.2. Problèmes de découpage des syntagmes

Les problèmes de **découpage** proprement dits se rencontrent lorsque, par exemple, on a des groupes prépositionnels (GP) successifs (ex : *le vin des sables du Golfe du Lion* ; *le vin de Bourgogne de l'épicier du coin* et *l'exposé de soutenance de thèse de mon frère* appellent trois types de parenthésages différents), une relative, un adjectif ou un GP à la suite d'un GN comportant un GP (ex : *Le grenier de la maison que Pierre a visitée* ; *le professeur de droit international* ; *le chien de la voisine sans mari* ne se structurent pas comme *le grenier de la maison que Pierre a visité* ; *le professeur de judo italien* ; *le chien de la voisine sans laisse*), un groupe prépositionnel à la suite d'un verbe et d'un complément nominal ou verbal (ex : *Marie a acheté une robe à fleurs* ; *Il a menacé de tuer sa voisine par balle* ne s'analysent pas comme *Marie a acheté une robe à Noël* ; *Il a menacé de tuer sa voisine par téléphone*), etc.

On voit que la pluralité possible de découpages d'une même suite d'unités peut conduire à des **ambiguïtés locales**. Si la suite est réellement ambiguë pour l'humain, le traitement devra construire les différentes structurations (ainsi pour *l'école de commerce de jeunes filles* ; *le professeur de football américain* ou *cette règle de la grammaire que Jean a écrite*). En revanche, dans les cas qui sont univoques pour l'humain (comme par exemple *la période de reconstruction de l'après-guerre* ou *la blouse de la vendeuse*

blonde), la pluralité des solutions n'est que le résultat "artefact" d'un traitement "non intelligent" : l'analyse syntaxique, pour se faire correctement, doit intégrer les informations nécessaires (par exemple sous forme de "traits sémantiques" spécifiant que l'adjectif *blond* peut qualifier un nom de type "humain" comme *vendeuse*, mais pas un nom de "vêtement" comme *blouse*) ; sur ce point, voir *infra* chapitre 5 § 1.1.

A titre d'illustration des problèmes de découpage et de structuration des groupes, nous proposons à la sagacité du lecteur la phrase suivante, relevée dans un hebdomadaire : *Cette entreprise a réalisé pour le Ministère une étude sur les besoins en docteurs de l'industrie qui apporte quelques éléments de réponse.*

Par ailleurs se posent également des problèmes de **portée** syntaxique de certains opérateurs (adverbes, quantificateurs, négation, circonstanciels, etc.), qui peuvent s'exprimer dans les mêmes termes que les précédents, mais qui ne leur sont pas assimilables. Ici aussi, il convient de distinguer les cas d'**ambiguïté** véritable, à décrire comme tels (ex : *Elle ne pleure pas parce qu'il est parti* : "elle pleure, mais ce n'est pas parce qu'il est parti, c'est pour une autre raison" ou "elle ne pleure pas : c'est parce qu'il est parti") des cas d'ambiguïté artefact, à éviter (ex : *Elle ne parle pas parce qu'elle est muette*, où la valeur sémantique de *muette* exclut la première interprétation), et des cas de double portée, dont nous avons parlé au §1.3.

2.3. Problèmes d'identification de fonction

Indépendamment des problèmes de découpage, l'identification des fonctions des constituants pose des difficultés spécifiques : c'est le cas par exemple pour des constructions où un constituant "manque" (ainsi le sujet de l'infinitive dans : *Il promet à Jean de venir* ou dans *Il permet à Jean de venir*), ou lorsqu'une fonction n'est pas marquée explicitement (ex : absence de préposition dans *Il chante le matin*), ou encore lorsque l'ordre des constituants n'est pas l'ordre canonique de la phrase noyau (ex : *Ce jour-là éclata un violent orage* ; *Il entend chanter l'hirondelle*). Pour traiter ces cas, des informations de toute nature sont nécessaires : informations syntaxiques (par exemple sur la différence entre *permettre* et *promettre* quant au contrôle du sujet de l'infinitive), mais aussi informations sémantiques et pragmatiques (par exemple pour différencier *j'entends chanter l'hirondelle* et *j'entends chanter la ritournelle*, ou *Il conduit les yeux fermés* et *Il conduit la voiture blindée*).

Là encore, des **ambiguïtés** sont possibles ; ainsi *La société a proposé à Jean de construire un analyseur* (qui le construit : Jean ou la société ?) ; *J'ai souvent vu manger des poulets* (les poulets sont-ils mangeurs ou mangés ?).

Au total, on le voit, les plus grosses difficultés pour les analyseurs syntaxiques concernent d'une part la structuration des constituants (en particulier

l'identification du point initial de rattachement, et de la fin d'une expansion) et d'autre part le calcul de la fonction des constituants lorsque celle-ci ne se laisse pas déduire directement de la position de ces constituants sur la chaîne. Bien entendu, ces difficultés sont considérablement accrues dans le cas des correcteurs orthographiques, et plus généralement des applications qui admettent des erreurs ou une information incomplète (typographie appauvrie, sans accents par exemple), puisque l'analyse ne peut plus utiliser de manière certaine l'information proprement morphologique.

3. Les formalismes syntaxiques

Nous ne proposerons ici qu'un rapide parcours, visant simplement à fixer quelques points de repère (que le lecteur pourra approfondir ultérieurement en consultant des ouvrages plus spécialisés) et à donner un aperçu de la diversité, ainsi que des limites, des formalismes existants.

3.1. Grammaires formelles (à règles de réécriture)

Rappelons brièvement qu'à la fin des années 50, les écrits de N. Chomsky (en particulier son article de 1956 et son ouvrage de 1957) ont marqué les débuts de la linguistique informatique. Rompant avec la tradition structuraliste distributionaliste, et introduisant une série de concepts nouveaux (notions de "compétence", de "récursivité", d'"acceptabilité syntaxique",...), il proposait de décrire la langue de façon formelle, par analogie avec les langages formels, en commençant par ce qui en constituait, à ses yeux, la pièce maîtresse : la syntaxe — l'objectif étant d'"engendrer" (de "générer") toutes les phrases grammaticales de la langue, et elles seules.

D'où la notion de **grammaire formelle**. Par définition, une grammaire formelle est la donnée d'un élément distingué, appelé **axiome** (ce sera ici P, qui symbolise la "phrase"), d'un ensemble de **règles de réécriture,** dites aussi **règles de production** (dont la forme la plus générale est : $u \longrightarrow v$), d'un **vocabulaire auxiliaire** (ce sont ici les symboles des catégories, comme N pour "nom", SN pour "syntagme nominal", etc.), et d'un **vocabulaire terminal** (ce sont les éléments du lexique).

Prenons l'exemple d'une grammaire extrêmement simple :

P —-> SN + SV
SN —-> art + N
SV —-> V + SN
V —-> *écrit, regarde*
N —-> *linguiste, grammaire*
art —-> *le, la.*

La flèche signifie que la partie gauche se réécrit sous la forme de la partie droite. Le symbole "+" symbolise l'opération de concaténation (ex : P se réécrit sous la forme d'un SN suivi d'un SV), tandis que la virgule marque un choix (article se réécrit soit *le* soit *la*).

Remarquons qu'une telle mini-grammaire correspond à une combinatoire déjà élevée et qu'elle peut engendrer d'une part des phrases agrammaticales (comme par exemple * *Le linguiste écrit le grammaire*) et d'autre part des phrases sémantiquement déviantes (comme *La grammaire écrit le linguiste* ; *La grammaire regarde le linguiste*, etc.).

A la suite de N. Chomsky, une **classification des grammaires** formelles a été élaborée, en fonction de la forme des règles qu'elles mettent en oeuvre (pour une présentation de cette classification, nous renvoyons à l'ouvrage de M. Gross et A. Lentin 1967). Cette typologie permet de distinguer les grammaires suivant leur "puissance d'expression", c'est-à-dire leur capacité à engendrer, avec un nombre fini de règles, un sous-ensemble infini plus ou moins contraint de phrases. Ainsi les grammaires dites "non contextuelles" (définies par le fait que le membre de gauche des règles est réduit à un seul élément, comme dans les règles ci-dessus) sont moins puissantes que les grammaires dites "contextuelles" (dans lesquelles un élément donné peut se réécrire différemment selon le "contexte", c'est-à-dire les éléments qui le suivent et / ou le précèdent sur la chaîne). Une des préoccupations majeures d'un certain nombre de théoriciens a été de limiter la puissance d'expression des grammaires utilisées pour décrire les langues (cf. T. Winograd, 1983, pp. 174 *sq*). Dans la pratique toutefois ce genre de considérations n'a qu'un intérêt très limité. D'une part, en toute rigueur, ces classifications n'ont de sens que pour des langages infinis (définis en termes de "compétences" théorique et non pas de "performance" observable), alors que les conditions même de l'analyse automatique (lexique fini et taille des phrases bornée) impliquent que l'ensemble des phrases analysables est fini. D'autre part, la réalité des analyseurs ne correspond pas à ces distinctions théoriques : en fait, aux règles de réécriture s'ajoutent toujours un certain nombre de mécanismes auxiliaires qui modifient considérablement la puissance d'expression de la grammaire ; ainsi, le plus souvent les règles de réécriture sont "non-contextuelles" mais des règles supplémentaires (comme les règles d'accord par exemple) donnent en fait à ces systèmes la puissance d'une grammaire contextuelle. On trouvera une présentation de l'état récent de cette question dans P. Sells *& al.* (eds) 1991 : en particulier on a pu montrer qu'un certain nombre de formalismes que nous évoquerons au § 3.3. *infra* font partie d'une même classe de grammaires qui ont été baptisées "grammaires modérément contextuelles" ("midly context - sensitive grammars").

3.2. Grammaires transformationnelles

Par-delà les grammaires à règles de réécriture (ou règles de dérivation) qui viennent d'être évoquées, se situent les grammaires transformationnelles. Formellement, une transformation consiste en l'application d'une ou plusieurs opérations élémentaires (de type suppression, ajout, déplacement d'un élément, ou substitution d'un élément à un autre), qui aboutit à modifier une structure syntagmatique et à la transformer en une autre.

Deux écoles linguistiques ont proposé le recours à des grammaires transformationnelles : l'école de N. Chomsky d'une part, celle de Z. Harris de l'autre.

Pour N. Chomsky, il s'agissait à l'origine de poser un niveau de structuration (dit "sous-jacent" ou "profond") où seraient spécifiées des relations et des fonctions syntaxiques non immédiatement appréhendables à partir des structures "de surface", et nécessaires pour une interprétation sémantique ultérieure ; les structures de surface étant reliées aux structures profondes par transformations. Dans ce cadre, des phrases superficiellement différentes mais syntaxiquement apparentées, comme par exemple *Il est facile d'implémenter cette grammaire* ; *Implémenter cette grammaire est facile,* et *Cette grammaire est facile à implémenter* reçoivent la même représentation profonde, et, à l'inverse, des phrases superficiellement similaires comme par exemple *La société a promis à Jean de construire un analyseur* et *La société a permis à Jean de construire un analyseur* ont des structures profondes différentes.

Les grammaires transformationnelles chomskiennes n'ont pas donné lieu à beaucoup de réalisations informatiques, d'une part à cause de la difficulté à formaliser et à implémenter la notion de transformation, d'autre part en raison des dérives successives de la théorie chomskienne depuis le modèle "standard" jusqu'à la plus récente "théorie du gouvernement et du liage" : à quelques exceptions près — citons par exemple les travaux réalisés sur l'anglais au Laboratoire d'Intelligence Artificielle du M.I.T. : cf. R. Berwick (1991), ainsi que l'analyseur de E. Wehrli sur le français : cf. E. Wehrli (1988) — la théorie chomskienne a été délaissée par la plupart des traitements automatiques, qui préfèrent rester dans le cadre des grammaires syntagmatiques.

Pour l'école de Z. Harris, les transformations opèrent entre des séquences de surface, reliant un sous-ensemble distingué des phrases de la langue (correspondant, schématiquement, aux phrases-noyaux élémentaires) à toutes les autres phrases. Ce cadre théorique a donné lieu à la construction d'analyseurs syntaxiques de l'anglais (cf. N. Sager 1981) et du français (cf. M. Salkoff 1973 et 1979), fondés sur des grammaires dites "en chaînes".

3.3. Quelques formalismes récents

Si quelques tentatives de traitement automatique ont été effectuées dans des cadres transformationnels, il reste que la plupart des analyseurs syntaxiques actuels se situent plutôt sur le terrain des grammaires à règles de dérivation évoqué au § 3.1. Dans cette perspective, toute une série de nouveaux modèles syntaxiques sont apparus depuis une vingtaine d'années.

Nous évoquerons tout d'abord le modèle de la "grammaire syntagmatique généralisée" ("generalized phrase structure grammar" : en abrégé GPSG) de G. Gazdar (cf. G. Gazdar *& al.*, 1985) et le modèle de la "grammaire syntagmatique guidée par la tête" ("head driven phrase structure grammar" : en abrégé HPSG) introduite par C. Pollard (cf. C. Pollard & I. Sag, 1987) ; ces deux modèles recourent au principe informatique de l'"unification" (voir *infra*, § 4.4.), d'où leur nom de "grammaires d'unification".

Les éléments de base de ces grammaires ne sont pas les catégories mais des "structures de traits" : si certaines de ces structures correspondent aux catégories classiques (ainsi, en GPSG, la structure {+N, -V, Barre : 2} représente un syntagme nominal complet), d'autres sont moins déterminées (la structure {+N} désigne un syntagme nominal ou adjectival, complet ou incomplet) et d'autres encore sont plus restrictives (la structure {+N, -V, Barre : 0, +masc, +sing} correspond à un nom masculin singulier). Les règles de la grammaire portent sur ces structures de traits, et permettent à la fois d'exprimer des conditions de nature diverse pour que la règle soit applicable, et de spécifier toutes sortes de caractéristiques (toujours sous forme de traits) du résultat obtenu. Il faut noter que les valeurs de certains traits peuvent être elles-mêmes des structures de traits, ce qui augmente considérablement la puissance du formalisme.

Dans le formalisme **GPSG**, les règles sont de plusieurs types :

- règles de "dominance immédiate", qui décrivent, sous forme de liste non ordonnée, les constituants composant un constituant de niveau hiérarchique immédiatement supérieur ;

- règles d' "ordre linéaire", qui donnent, indépendamment des règles précédentes, des conditions nécessaires portant sur l'ordre des constituants sur la chaîne linéaire ;

- règles de "restriction sur la co-occurrence des traits" et de "spécification de traits par défaut", qui s'ajoutent à quelques principes généraux (comme le "principe des traits de tête" et le "principe du contrôle et de l'accord") pour décrire diverses contraintes (sur la constitution de nouvelles structures de traits et sur les valeurs de ces traits) dans l'application des règles de dominance.

Le modèle **HPSG** est une variante de GPSG qui limite le nombre de règles en encodant plus d'information combinatoire dans le lexique, en parti-

culier les constructions verbales. Moins contraint et plus souple que GPSG, mieux adapté aussi à l'unification, ce modèle connaît aujourd'hui un succès croissant.

La "Grammaire lexicale fonctionnelle" ("Lexical Functional Grammar" en abrégé **LFG**) développée par J. Bresnan (cf. R. Kaplan & J. Bresnan 1982) est aussi une "grammaire d'unification", mais dont les principes sont radicalement différents. L'information sur les constructions syntaxiques est, pour l'essentiel, directement codée dans le lexique, où elle est décrite en termes de fonctions (sujet, objet direct, etc.) et non pas en termes de catégories. Il existe tout de même des règles de dérivation, mais elles sont très générales et non contraintes : elles ne fournissent en somme qu'un cadre dans lequel va pouvoir s'opérer l'analyse, guidée et filtrée par les informations lexicales. Celle-ci conduit à une double représentation : une "c-structure" qui décrit la décomposition en syntagmes, et une "f-structure", qui décrit les relations fonctionnelles sur lesquelles reposent toutes les contraintes qui ont dû être satisfaites pour aboutir à l'analyse correcte.

Evoquons également la théorie des grammaires **catégorielles**, qui a été introduite par Y. Bar-Hillel (1964), et a été utilisée en particulier par R. Montague (1970) — cf. *infra* chapitre 5. Cette théorie a connu des développements récents avec notamment la "grammaire d'unification catégorielle" de H. Uszkoreit (1986) et la "grammaire catégorielle généralisée" de M. Moortgat (1989). Dans ce type de grammaires, on définit un nombre très restreint de catégories dites "de base" (typiquement P la phrase, N le nom, et les syntagmes SN et SP). Les autres catégories sont obtenues par des combinaisons plus ou moins complexes des catégories de base, encodant ainsi dans leur symbolisme les règles de la grammaire. Ainsi la catégorie "déterminant" sera codée SN/N parce qu'elle forme un SN quand elle est combinée à droite avec un N. De même, à un verbe intransitif sera attribuée la catégorie P\SN (combiné avec un SN à gauche, il forme une phrase) et un verbe transitif s'écrira (P\SN)/SN (combiné à droite avec un SN, il donne un syntagme verbal P\SN, qui doit être à son tour combiné avec un SN à gauche pour former une phrase). Les règles de dérivation n'ont donc pas à être explicitées dans ce formalisme. Elles sont en fait issues des deux seules règles "d'application fonctionnelle" :

$$X \longrightarrow X/Y\ Y$$

$$X \longrightarrow Y\ X\backslash Y$$

où X et Y sont des catégories quelconques (complexes ou de base).

Il faut ajouter à ce principe de base un certain nombre de mécanismes (comme la "composition de fonctions" et "la montée de type") pour augmenter la puissance d'expression de ce formalisme, qui se heurte vite à des difficultés de représentation peu compatibles avec l'analyse de phrases tout-venant.

Bien d'autres formalismes ont été proposés ; on peut par exemple citer les "grammaires d'arbres adjoints" ("Tree Adjoining Grammars" ou **TAGs**)

de A. Joshi (1987), dont les éléments de base sont des arbres partiels sur lesquels est définie une opération "d'adjonction" : les règles de réécriture opèrent sur des arbres élémentaires, l'adjonction remplaçant la concaténation.

3.4. Limites des formalismes existants

Au-delà de leur diversité, tous ces formalismes présentent un point commun : ils imposent chacun un système formel de description, caractérisé par un petit nombre d'opérateurs agissant sur des représentations homogènes. Ce souci d'économie est bien sûr louable, et il répond à deux préoccupations dont l'importance est indiscutable : d'une part, au plan linguistique, la mise en évidence d'un petit nombre de mécanismes de base qui expliquent les régularités fondamentales de la syntaxe d'une langue, et d'autre part, au plan informatique, la mise en place d'un langage de description suffisamment précis pour pouvoir écrire des analyseurs dont la taille et la complexité ne sont pas conciliables avec une programmation faite de "bricolages".

On peut se demander néanmoins si ce type d'approche pourra un jour déboucher sur des analyseurs capables de traiter des textes tout-venant. A l'heure actuelle, ces formalismes ne décrivent facilement qu'un petit sous-ensemble des phrases d'une langue comme le français : essentiellement des phrases "canoniques", à construction très régulière, qui, comme nous avons essayé de le montrer (cf. *supra*, § 1. et § 2.), ne sont pas réellement représentatives des problèmes à prendre en compte (cf. A. Abeillé (ed.) 1991). En effet, ce qui caractérise ces constructions canoniques, c'est une biunivocité entre fonction, catégorie et position : ainsi le sujet est un syntagme nominal, et ce syntagme nominal est celui qui est antéposé au verbe. Or dans les textes, toutes sortes d'autres cas se produisent : le sujet peut ne pas être un GN, il peut ne pas être antéposé, etc. On peut simplement énoncer un principe de "naturalité" du genre "le sujet a tendance à être un GN et à être antéposé au verbe". Le dilemme pour les formalismes syntaxiques est alors le suivant : soit on se limite aux constructions les plus régulières et on doit renoncer à une large couverture, soit au contraire on décrit dans la grammaire tous les cas de figure et alors, comme le démontre avec force J.P. Chanod (1992), on génère énormément de solutions "parasites" sur les phrases les plus banales, puisque toutes les règles (les "naturelles" et celles qui le sont moins) sont mises sur le même plan. On peut donc partager le scepticisme de F. Segond qui se demande : "existe-t-il, actuellement, un seul formalisme qui puisse être un outil de construction d'un analyseur morpho-syntaxique à large couverture ?" (F. Segond, 1991, p.25). Bien sûr, pour chacune de ces théories, des efforts considérables sont actuellement déployés pour étendre la liste des phénomènes couverts (souvent d'ailleurs au prix de quelques "entorses" à l'unicité du formalisme). Mais il n'est pas prouvé que cela suffira, étant donné la diversité des difficultés non encore résolues : le risque que les extensions successives ne finissent par dissoudre le formalisme initial est tout à fait

réel. Il faut aussi noter que la recherche dans ce domaine ne se donne pas forcément comme objectif l'obtention d'analyseurs à large couverture, mais s'intéresse plutôt à des problèmes théoriques : c'est ce qui explique, par exemple, l'importance donnée par ces recherches à quelques phénomènes linguistiques très précis dont l'intérêt théorique est indéniable, mais qui sont loin d'être les plus "urgents" pour des systèmes de traitement automatique en grandeur réelle. C'est le cas par exemple des "dépendances non bornées" (à l'oeuvre dans des constructions — plus fréquentes d'ailleurs en anglais qu'en français — du type : *A qui croyez-vous que Jean pensait qu'il fallait s'adresser ?* où le syntagme *à qui*, qui apparaît dans la principale, est à rattacher à une subordonnée à un niveau hiérarchique non déterminable à l'avance).

Il faut signaler que parallèlement à ces travaux, des approches moins théoricistes voient le jour : sans tomber dans le travers du "bricolage" que nous dénoncions, certains chercheurs se donnent comme unique objectif la tâche informatique d'analyse automatique. Ils sont bien sûr amenés à construire des systèmes de règles qui ne sont pas sans rapport avec celles utilisées dans les formalismes que nous avons présentés, mais comme ils ne visent pas un niveau théorique de description, ils gagnent en souplesse de traitement grâce à une certaine hétérogénéité théorique. L'envers de la médaille est évident : on n'a pas la certitude que l'extension progressive de ces systèmes ne conduira pas à des incohérences. Là non plus le pari n'est pas gagné d'avance... On peut citer comme exemple de ce type d'approche les travaux de l'équipe de J.P. Chanod à IBM-France (cf. J.P. Chanod 1991 et S. Guillemin-Lanne 1991) et ceux d'une PME québécoise : Machina Sapiens (cf. Cl. Coulombe 1991). Ces deux projets d'analyseurs syntaxiques à grande couverture sont sans aucun doute parmi les plus avancés aujourd'hui pour le français, et dans les deux cas, ce n'est sans doute pas un hasard, un des objectifs essentiels poursuivis est la réalisation d'un correcteur orthographique.

4. Les analyseurs syntaxique : principes

Pour une présentation succinte des principaux analyseurs syntaxiques "historiques", classés selon les types de grammaires formelles évoquées plus haut (cf. *supra*, § 3.1.), nous renvoyons le lecteur à G. Sabah (1989 : ch.2, § 2.3.) ; pour une présentation de quelques systèmes récents d'analyse syntaxique du français, voir Ch. Fay-Varnier *& al.* (1991).

Nous considérerons successivement la représentation des connaissances syntaxiques, les stratégies d'analyse effective de la phrase qu'ils mettent en oeuvre, les stratégies de traitement des ambiguïtés locales auxquelles ils recourent, et enfin les techniques strictement informatiques sur lesquelles certains d'entre eux reposent.

4.1. Représentation des connaissances syntaxiques

La quantité de connaissances dont l'analyseur doit disposer est considérable : cela impose donc d'utiliser des techniques de représentation qui permettent de les entrer, de les modifier et de les compléter facilement. Aujourd'hui, pratiquement tous les systèmes utilisent pour cela une méthode **déclarative**. Les connaissances sont décrites dans des fichiers-texte, dans des formats "lisibles" par un humain, qui peut aisément les contrôler et les modifer (sans forcément posséder des compétences informatiques de haut niveau). Ces connaissances sont ensuite utilisées par le système d'analyse, soit directement, par un **interpréteur** (capable de "lire" ce format), soit après avoir été transformées par un **compilateur** (qui traduit ce format dans un autre, plus adapté aux traitements par la machine, mais illisible pour l'humain). L'avantage de la compilation est d'augmenter la vitesse de traitement. L'avantage des interpréteurs est de permettre des modifications des connaissances de manière interactive, au cours même d'une session d'analyse. Lors de la mise au point d'une grammaire, ces derniers sont donc préférables ; en revanche, une fois opérationnel, un système compilé sera plus performant.

Les connaissances syntaxiques sont de deux sortes : des **règles** d'une part, et des informations attachées au **lexique** de l'autre. La nature et l'importance respective de ces deux types de connaissances varient énormément d'un système à l'autre : on a vu que les théories formelles divergeaient beaucoup à cet égard. Alors que les dictionnaires morphologiques contiennent à peu près les mêmes informations, les dictionnaires à vocation syntaxique sont très différents les uns des autres et intimement liés au système pour lequel ils ont été conçus : cela constitue un frein au développement d'outils généraux. Il faut noter cependant l'entreprise de l'équipe du LADL, qui construit depuis de nombreuses années des "lexiques-grammaires" de diverses langues (à commencer par le français) : il s'agit là sans doute de l'outil le plus complet dans ce domaine, en particulier en ce qui concerne les constructions verbales (pour une bibliographie des recherches sur les lexiques-grammaires, cf. C. Leclère & C. Subirats-Rüggeberg 1991).

Rappelons enfin que beaucoup de systèmes traitent en même temps l'analyse morphologique et l'analyse syntaxique, ce qui n'est pas sans répercussions sur l'organisation des lexiques associés à ces systèmes. De même, beaucoup de systèmes incorporent des connaissances sémantiques, aussi bien dans le lexique que dans les règles, soit que ces connaissances soient utilisées pour l'analyse syntaxique elle-même (nous avons vu de nombreux exemples où ce type d'information semble irremplaçable), soit que le système intègre dans un même module analyse syntaxique et analyse sémantique.

4.2. Stratégies d'analyse

Pour procéder à une **analyse effective** de la phrase, il faut appliquer à cette phrase les règles syntaxiques proposées par la description linguistique. Classiquement, pour un système de règles dérivationnelles, il existe deux grands types de stratégies informatiques, nommées respectivement "analyse descendante" et "analyse montante" (pour une présentation détaillée, voir T. Winograd 1983, § 3.4., pp.93-108).

L'analyse **descendante**, dite aussi "analyse dirigée par les hypothèses", consiste à aller des buts (les structures données par les règles) aux faits (les mots de la phrase). Partant de l'axiome P elle compare à chaque étape avec la phrase à analyser : s'il n'y a pas coïncidence, on applique une règle de réécriture sur l'élément le plus à gauche, et s'il y a coïncidence sur les deux éléments les plus à gauche (dans la phrase et dans la suite de symboles engendrée par la ou les règle(s) déjà appliquée (s)), on les élimine et on continue. Ainsi, étant donné les six règles présentées plus haut (§ 3.1.), et la phrase *Le linguiste écrit la grammaire*, une analyse descendante suivrait les étapes suivantes (dans l'hypothèse d'une lecture linéaire de la phrase de gauche à droite, qui est le type de lecture le plus couramment utilisé par les analyses descendantes) :

a) *le linguiste écrit la grammaire* : comparaison avec "P" : pas de coïncidence : application de la règle P —-> SN + SV ; d'où suite = "SN + SV"

b) *le linguiste écrit la grammaire* : comparaison avec "SN + SV" : pas de coïncidence : application de la règle SN —-> art + N ; d'où suite = "art + N + SV"

c) *le linguiste écrit la grammaire* : comparaison avec "art" : pas de coïncidence : application de la règle art —-> *le* ; d'où suite = "*le* + N + SV"

d) *le linguiste écrit la grammaire* : comparaison avec "*le*" : coïncidence : élimination de "*le*" ; d'où suite = "N + SV"

e) *linguiste écrit la grammaire* : comparaison avec "N + SV" : pas de coïncidence : application de la règle N —-> *linguiste* ; d'où suite = "*linguiste* + SV"

f) *linguiste écrit la grammaire* : comparaison avec "*linguiste* + SV" : coïncidence : élimination de "*linguiste*" ; d'où suite = "SV"

g) *écrit la grammaire* : comparaison avec "SV" : pas de coïncidence : application de la règle SV —-> V + SN ; d'où suite = "V + SN"

h) *écrit la grammaire* : comparaison avec "V + SN" : pas de coïncidence : application de la règle V —-> *écrit* ; d'où suite = "*écrit* + SN"

i) *écrit la grammaire* : comparaison avec "*écrit* + SN" : coïncidence : élimination de "*écrit*" ; d'où suite = "SN"

j) *la grammaire* : comparaison avec "SN" : pas de cioïncidence : application de la règle SN —-> art + N ; d'où suite = "art + N"

k) *la grammaire* : comparaison avec "art + N" : pas de coïncidence : application de la règle art —-> *la* ; d'où suite = "*la* + N"

l) *la grammaire* : comparaison avec "*la* + N" : coïncidence : élimination de "*la*" ; d'où suite = "N"

m) *grammaire* : comparaison avec "N" : pas de coïncidence : application de la règle N —-> *grammaire* ; d'où suite = "*grammaire*"

n) *grammaire* : comparaison avec "*grammaire*" : coïncidence : élimination de "*grammaire*" ; d'où suite = "O"

o) *O* : comparaison avec "O" : fin de l'analyse.

Les avantages et les inconvénients de l'analyse descendante sont clairs : d'un côté elle évite de construire des constituants qui, en fin d'analyse, se révéleraient impossibles à rattacher, et elle permet de ne pas retenir des catégories non pertinentes dans un environnement donné en cas de polycatégorie d'un élément (la prise en compte des règles de réécriture du SN permet par exemple de voir tout de suite qu'à la gauche d'un N, *la* ne peut pas être un pronom personnel mais seulement un article), d'un autre côté une analyse purement descendante risque d'effectuer de mauvais rattachements d'ajouts qu'un simple "regard en avant" permettrait d'éviter (ainsi dans *la règle du système que le linguiste a écrite*, il suffirait de pouvoir considérer l'accord du participe pour éliminer le rattachement erroné du relatif au N situé au plus près (*le système*) et le rattacher au bon antécédent *la règle*).

L'analyse **ascendante**, dite aussi "dirigée par les données", consiste à l'inverse à aller des faits (mots de la phrase) vers les buts (structures). Au fur et à mesure de la lecture des mots de la phrase, elle tente de remplacer chaque mot par sa (ou l'une de ses) catégorie(s), puis elle parcourt les règles "à l'envers" pour réécrire par étape les éléments obtenus en constituants de plus en plus larges, jusqu'à remonter à l'axiome P. Ainsi, pour analyser la même phrase que précédemment, avec la même grammaire, on aurait les étapes suivantes (dans l'hypothèse d'une lecture de gauche à droite procédant par choix d'éléments de taille croissante, assez couramment utilisé par les analyses montantes) :

a) *le linguiste écrit la grammaire* : lecture de "*le*" et application de la règle art —-> *le* (à l'envers, c'est-à-dire réécriture de "*le*" sous la forme "art")

b) art *linguiste écrit la grammaire* : lecture de "*linguiste*" et appplication de la règle N —-> *linguiste* (réécriture de "*linguiste*" sous la forme "N")

c) art N *écrit la grammaire* : lecture de "art N" et application de la règle SN —-> art + N (réécriture de "art N" sous la forme "SN")

d) SN *écrit la grammaire* : lecture de "*écrit*" et application de la règle V —-> *écrit* (réécriture de *écrit* sous la forme "V")

e) SN V *la grammaire* : lecture de "*la*" et application de la règle art —-> *la* (réécriture de "*la*" sous la forme "art")

f) SN V art *grammaire* : lecture de "*grammaire*" et application de la règle N —-> *grammaire* (réécriture de "*grammaire*" sous la forme "N")

g) SN V art N : lecture de "art N" et application de la règle SN —-> art + N (réécriture de "art N" sous la forme "SN")

h) SN V SN : lecture de "V SN" et application de la règle SV —-> V + SN (réécriture de "V SN" sous la forme "SV")

i) SN SV : lecture de "SN SV" et application de la règle P —-> SN + SV (réécriture de "SN SV" sous la forme "P"

j) P : fin de l'analyse.

Là encore, les avantages et les inconvénients de ce type d'analyse apparaissent clairement : d'un côté l'analyse montante est plus flexible en cas d'erreur de construction, mais d'un autre côté, elle produit elle aussi certaines ambiguïtés artificielles (en particulier en cas de polycatégorie d'un mot).

Bien que les deux types d'analyse soient formellement équivalents, on voit qu'ils se heurtent à des difficultés de nature différente. Diverses stratégies **mixtes** ont donc été mises au point, qui cherchent à pallier les inconvénients de ces deux méthodes. Par exemple, certains systèmes utilisent une stratégie guidée partiellement par les données : certaines données servent à déclencher des règles, qui, une fois sélectionnées, s'appliquent de manière descendante. Cela a l'avantage d'utiliser au mieux le fait que, dans une langue comme le français, les débuts de syntagmes sont souvent introduits par des marqueurs spécifiques ; cela permet aussi d'utiliser facilement des règles en partie lexicales, comme les constructions verbales ou adjectivales.

Les exemples que nous venons de donner supposent que l'on a choisi de traiter la phrase de gauche à droite, en commençant par le premier mot. Rien n'interdit de procéder autrement. On peut choisir un **sens de lecture** droite-gauche, et aussi de ne pas traiter les mots systématiquement dans l'ordre où ils se présentent : les stratégies ascendantes utilisent souvent la technique dite des "îlots de confiance", dans laquelle on cherche à analyser des segments non ambigus les plus longs possibles, quelle que soit leur position dans la phrase, et l'on étend ensuite à partir de ces groupes (aussi bien à droite qu'à gauche) jusqu'à l'analyse complète de la phrase. Cette technique est particulièrement intéressante pour les traitements de données incertaines (en particulier les correcteurs orthographiques). Signalons aussi que certains systèmes permettent à l'utilisateur de choisir lui-même en partie la stratégie d'analyse. Ainsi le système LEU/2, développé dans le cadre du projet LILOG (cf. O. Herzog & C.R. Rollinger (eds.), 1991, 74-87), réalise une analyse ascendante avec de nombreuses options (gauche-droite, droite-gauche, "profondeur d'abord" : c'est-à-dire priorité à l'extension de sous-arbres déjà construits, "largeur d'abord" : c'est-à-dire priorité aux unités non encore traitées, priorité aux chaînes les plus longues, priorité aux règles les plus fréquentes, etc.). Cette possibilité de choix d'une stratégie est très liée à une technique qui

connaît un succès croissant : il s'agit de la technique dite des "**charts**" (cf. J. Kaplan 1979, et T. Winograd 1983 pp. 120-128) qui consiste à conserver en mémoire, sous une forme appropriée, les constructions partielles déjà effectuées (qui sont donc faites une fois pour toutes). Cette représentation très "déclarative" de l'état d'avancement de l'analyse se prête bien à un contrôle du processus très explicite : à tout instant, les différentes tâches à accomplir sont consignées dans ce que l'on appelle un "agenda", et la sélection de la tâche qui va être exécutée peut obéir à n'importe quels critères précisés à l'avance.

4.3. Traitement des ambiguïtés locales

Dans les cas où un analyseur se trouve devoir faire un choix entre plusieurs solutions possibles et n'est pas en mesure de trancher (situation **d'ambiguïté** effective ou artificielle), plusieurs stratégies sont possibles : les deux stratégies les plus classiques sont le "retour en arrière" et le "parallélisme".

La stratégie du **retour en arrière** consiste à choisir l'une des solutions (de façon plus ou moins arbitraire selon les cas : on peut échapper à l'arbitraire total en se donnant un ordre de plausibilité des différentes solutions, ou une heuristique de choix) et à continuer l'analyse en mémorisant les solutions non retenues et l'état de l'analyse au moment du choix. Si la suite de l'analyse remet en cause le choix effectué, alors on retourne à l'endroit du choix et l'analyse repart en sélectionnant une autre solution. Si le premier choix est le bon, cette stratégie est payante car elle gagne beaucoup de temps de traitement, par contre s'il faut effectuer plusieurs retours en arrière, le temps de traitement est long (car cela conduit souvent à refaire plusieurs fois certains calculs) ; par ailleurs les cas d'ambiguïté effective ne sont pas dépistés : dès lors que le premier choix marche, aucune autre solution n'est essayée.

La stratégie inverse du **parallélisme** consiste à développer toutes les solutions possibles lorsqu'un choix se présente, et donc à effectuer toutes les analyses possibles de la phrase. Si elle présente l'inconvénient d'impliquer des temps de calcul relativement longs et d'entraîner des difficultés techniques liées à l'écriture d'algorithmes simulant le parallélisme, en revanche cette stratégie permet de ne pas refaire plusieurs fois les mêmes calculs, et de donner toutes les solutions dans les cas d'ambiguïté globale effective.

Là encore, des stratégies alternatives ont été développées pour éviter ce dilemme. Certaines se donnent pour objectif d'effectuer les choix pertinents au moment approprié, en évitant aussi bien les solutions parasites que la réitération de calculs identiques : ce sont des analyseurs **déterministes**. Grâce à une technique de "regard en avant", les choix peuvent s'appuyer sur la présence d'éléments qui suivent l'élément sur lequel porte l'ambiguïté ; cf. M. Marcus (1980) ou encore le système ANDI du LIMSI (cf. G. Sabah 1989

chapitre 5). Avec l'apparition des **"charts"** (cf. *supra*, § 4.2.), le traitement des ambiguïtés peut être contrôlé de manière beaucoup plus souple. Cette technique permet en effet d'implémenter une série de règles spécialement dédiées à ce problème : le système peut du coup réagir différemment suivant le type d'ambiguïté. Enfin signalons que se développent (en particulier pour des analyseurs à large couverture) des **algorithmes spécifiques**, conçus spécialement pour éviter la prolifération de solutions parasites. Ainsi FROG, l'analyseur du correcteur orthographique d'IBM-France, effectue une analyse en deux temps : une première passe conduit à une "esquisse syntaxique", étape préliminaire qui identifie en gros les syntagmes minimaux ; ensuite une deuxième passe s'attaque aux problèmes plus difficiles, générateurs d'ambiguïtés massives, comme le rattachement des syntagmes prépositionnels, les constructions verbales, etc. (cf. J.P. Chanod 1991).

4.4. Les outils de traitement

Les premiers outils de traitement suffisamment systématiques et généraux conçus pour procéder à l'analyse syntaxique ont été les "réseaux de transitions enrichis" (Augmented Transition Networks —en abrégé ATN, introduits par W. Woods, 1970). Les **ATN** permettent de traduire de manière assez directe des règles de dérivation en un mécanisme procédural dans lequel, en gros, l'application d'une règle correspond à passer d'un état à un autre en suivant un "chemin" dans un réseau (pour une description détaillée des ATN, voir T. Winograd, 1983, chapitre 5 ou G. Sabah, 1989, chapitre 3), l'arbre syntaxique se construisant au fur et à mesure. Un système de "conditions-actions" confère à cet outil une puissance suffisante pour pouvoir conserver n'importe quel type d'information au cours du traitement et pour pouvoir exécuter chaque fois que c'est nécessaire des actions spécifiques liées à ces informations.

Les ATN ne sont de fait plus guère utilisés depuis quelques années. On leur reproche de ne pas être suffisamment déclaratifs : le système de conditions-actions conduit à écrire une grande partie de l'expertise linguistique sous forme procédurale, ce qui le rend assez lourd à gérer.

L'outil le plus populaire aujourd'hui est sans conteste l'**unification**. Son succès s'explique à la fois par son adéquation à la plupart des grammaires formelles (GPSG, HPSG, LFG, etc.) et par sa facilité d'implémentation dans des langages de programmation modernes que l'on appelle les **langages à objets**.

Il n'est pas question dans le cadre de cet ouvrage d'exposer les différents formalismes d'unification. Nous nous contenterons d'en donner le principe de base. On décrit les différents éléments de la grammaire sous forme d'objets qui sont des listes d'attributs-valeurs (ou structures de traits), sur lesquels on définit une opération, l'unification, qui consiste à construire un nouvel objet fusionnant les listes des opérandes, quand cela est possible,

c'est-à-dire quand les valeurs des attributs sont compatibles. Prenons un exemple très simpliste. Soient les objets suivants :

O1 = { lex = *le* ;
cat = DET ;
genre = masc ;
nombre = sing }

O2 = { lex = *linguiste* ;
cat = N ;
nombre = sing }

O3 = { cat = GN ;
schéma = (x_1, x_2) ;
x_1 = { cat = DET } ;
x_2 = { cat = N } ;
genre = <x_1 genre> = <x_2 genre> ;
nombre = <x_1 nombre > = <x_2 nombre > }

O1 et O2 pourraient être des représentations (partielles) des éléments lexicaux *le* et *linguiste*, tandis que O3 serait une représentation (simplifiée, elle aussi) de la règle GN —> DET + N. L'unification de ces objets revient à appliquer cette règle en construisant un quatrième objet O4 de la forme :

O4 = { lex = *le linguiste* ;
cat = GN ;
schéma = (x_1, x_2) ;
x_1 = { lex = *le* ;
cat = DET ;
genre = masc ;
nombre = sing } ;
x_2 = { lex = *linguiste* ;
cat = N ;
genre = masc ;
nombre = sing } ;
genre = masc ;
nombre = sing }

L'unification a été possible dans ce cas parce qu'un certain nombre de conditions exprimées dans O3 (sur les catégories de x_1 et x_2, sur les accords de genre et de nombre) étaient réalisées. De plus l'unification a permis de compléter des informations manquantes (ainsi le genre de *linguiste*, qui était lexicalement indéterminé a pris la valeur masculin dans O4). On le voit : un des intérêts majeurs de l'unification est de permettre de décrire des régularités de la langue sans se préoccuper de savoir si elles seront utilisées comme conditions de déclenchement d'une règle ou pour déterminer la nature exacte d'un constituant. Notons aussi que des informations de tout type (en particulier sémantique) peuvent être codées dans ce format attribut-valeur.

Il existe de nombreux systèmes informatiques implémentant une variante ou une autre de ce principe d'unification. Citons le système de M. Kay (1979) appelé "Functional unification grammar", le système PATR-II de S. Shieber *& al.* (1983), le système STUF ("Stuttgard Type Unification Formalism" utilisé dans LILOG). On peut aussi placer dans cette catégorie les DCG ("Definite Clause Grammars") qui utilisent directement le mécanisme d'unification inhérent au langage de programmation PROLOG. Il est important de noter que ces systèmes ne définissent pas des formalismes grammaticaux spécifiques, contrairement à ce que l'appellation de certains d'entre eux pourraient laisser croire. En fait, il s'agit réellement d'outils qui peuvent être utilisés dans le cadre de différentes théories syntaxiques : ainsi STUF a servi à implémenter aussi bien une grammaire catégorielle qu'une grammaire HPSG.

Tous les analyseurs existants ne sont pas pour autant fondés sur ce principe de l'unification. Il existe bien d'autres techniques, souvent d'ailleurs elles aussi décrites dans des langages à objets, avec des représentations attributs-valeurs, mais dans lesquelles le "moteur" de traitement n'est pas réduit à une opération unique comme l'unification. En particulier se développent beaucoup aujourd'hui des traitements plus **distribués**, profitant de la possibilité "d'encapsuler" des traitements dans les objets. De même sont de plus en plus utilisées des architectures, comme les "**tableaux noirs**", qui permettent de faire **coopérer** des modules de traitement disposant d'expertises complémentaires. En effet une architecture à base de "tableau noir" (cf. R. Engelmore & T. Morgan (eds) 1988) est composée de trois types de structures distinctes : d'abord, des "experts" indépendants les uns des autres, qui sont autant de spécialistes dotés de connaissances spécifiques (par exemple, on peut envisager un expert "morphologique", un expert des constructions verbales, etc.) ; ensuite le tableau noir proprement dit, qui est une structure de données dans laquelle on inscrit au fur et à mesure les résultats partiels obtenus et les tâches à accomplir (on le voit, cette structure est une généralisation des "agendas" mis en oeuvre dans la technique des "charts", cf. *supra*, § 4.2.) ; enfin un contrôleur, qui est en fait lui aussi un expert, dont le rôle est de sélectionner à tout moment le spécialiste le mieux à même d'avancer dans la résolution du problème, en fonction de l'état actuel du tableau noir. On conçoit aisément tout l'intérêt de ce type d'architecture pour permettre la prise en compte simultanée de critères de différents niveaux (morphologiques, lexicaux, sémantiques, etc.).

5. Perspectives

A l'heure actuelle, les tentatives visant à améliorer la qualité et les performances des analyseurs syntaxiques se situent sur deux plans :

- d'une part la recherche d'**architectures informatiques** plus efficaces : en particulier dans le domaine de l'intelligence artificielle, les chercheurs travaillent à élaborer des systèmes où les diverses sources de connaissances, exprimées sous forme déclarative, seraient susceptibles d'**interagir** (coopération de sources indépendantes, comme dans la technique du "tableau noir", ou contrôle distribué dans divers modules qui peuvent s'influencer mutuellement, ou encore architectures non modulaires), et à mettre au point des systèmes capables d'effectuer des calculs en parallèle ; en particulier, l'idée de pondérer les règles, pour mieux gérer leur interaction et rendre compte du principe de "naturalité" (cf. *supra*, § 3.4.), commence à prendre de l'importance. Dans le même ordre d'idées, les techniques **connexionnistes** (qui permettent d'implémenter des mécanismes d'apprentissage et de pondération) ont fait une timide apparition en analyse syntaxique, essentiellement pour traiter des phénomènes spécifiques, comme le rattachement de groupes prépositionnels (cf. par exemple J.L. Mc. Clelland & A.H. Kawamoto 1986, ou S. Wermter & W. Lehnert 1989).

- d'autre part la recherche de **formalismes grammaticaux** permettant d'écrire des grammaires à large couverture, capables de gérer des phénomènes linguistiques complexes (en particulier une pluralité de sources d'ambiguïtés syntaxiques différentes) et de proposer de meilleures interactions entre syntaxe et sémantique.

Paradoxalement, dans les recherches sur les analyseurs syntaxiques, tout se passe à l'heure actuelle comme si la description **linguistique** des phénomènes syntaxiques allait de soi, les débats ne portant que sur la nature des formalismes susceptibles de représenter ces phénomènes, et sur les techniques d'implémentation. Pourtant, même sur des langues aussi massivement décrites que l'anglais ou le français, bien des points de syntaxe demeurent mal connus et mal décrits, et l'on peut se demander dans quelle mesure l'acharnement à construire des formalisations de plus en plus sophistiquées ne conduit pas à une sorte de cécité face à la multitude et à la complexité des phénomènes à observer et à décrire.

Catherine FUCHS et Bernard VICTORRI
(ELSAP-CNRS)

Repères bibliographiques

1. Présentations d'ensemble :

CARRE, R. *& al.* (1991) : *Langage humain et machine*, Paris, Presses du CNRS.

[Le ch. 1 : "L'étude et le traitement automatique des langues naturelles" et le ch. 3 : "Le traitement de l'écrit" contiennent, entre autres, quelques développements sur les questions d'analyse syntaxique automatique.]

COULON, D. & KAYSER, D. (1986) : Informatique et langage naturel : présentation générale des méthodes d'interprétation des textes écrits, *Technique et Science Informatiques*, 5 : 2, Paris, Gauthier-Villars, 103-128.

[Article de synthèse ; sur l'analyse syntaxique, voir pp. 108-111.]

KAYSER, D. (1985) : Des machines qui comprennent notre langue, *La Recherche*, 16 : 170, Paris, 1198-1212.

[Article de vulgarisation consacré aux problèmes de la compréhension automatique ; sur l'analyse syntaxique, voir pp. 1202 *sq.*]

SABAH, G. (1988) et (1989) : *L'intelligence artificielle et le langage* ; vol. I : "Représentations des connaissances" ; vol. II : "Processus de compréhension", Paris, Hermès.

[Voir notamment le vol. I, ch.2 : présentation des grammaires formelles ; et les pp. 143-145 du ch. 5 : présentation des "lexiques-grammaires" de M. Gross.]

[Voir également le vol. II, ch. 2 : "Les principes de l'analyse des phrases, ch. 3 : "Réseaux de transitions", ch. 4 : "Analyseurs dirigés par le lexique" et ch. 5 : "Analyseurs déterministes".]

SMITH, G. (ed.) (1991) : *Computers and human language*, San Mateo, Kaufmann.

[Voir le ch. 6 : "Approaches to syntax" et le ch. 7 : "Augmented parsers and modern grammars".]

WINOGRAD, T. (1983) : *Language as a cognitive process*, (vol.1 : "Syntax"), Reading, Mass., Addison Wesley.

[Ouvrage de référence pour une présentation synthétique et détaillée des méthodes d'analyse syntaxique utilisées en traitement automatique.]

2. Modèles linguistiques et formalismes de représentation syntaxique :

BAR-HILLEL, Y. (1964) : *Language and Information*, Reading, Mass., Addison-Wesley.

[Une bonne introduction à la grammaire catégorielle.]

CHANOD, J.P. (1992) : Dérèglement du langage et parasitisme computationnel : problèmes de robustesse en analyse syntaxique, *T.A. Informations*, Paris, Klincksieck, (à par.).

[Article décrivant de manière convaincante les difficultés inhérentes aux formalismes classiques pour traiter des textes tout venant.]

CHOMSKY, N. (1957) : *Syntactic Structures*, La Haye, Mouton ; trad.fr. (1969) *Structures syntaxiques*, Paris, Le Seuil.

[Ouvrage classique présentant la "grammaire générative-transformationnelle" chomskienne (première version) : introduction des notions de grammaire formelle, de grammaires syntagmatiques, et de transformation.]

CHOMSKY, N. (1975) : *Aspects of the theory of syntax*, Cambridge, Mass., M.I.T. Press ; trad. fr. (1971) : *Aspects de la théorie syntaxique*, Paris, Le Seuil.

[Ouvrage de référence de la version dite "standard" de la grammaire générative-transformationelle.]

CLEMENT, D. (éd.) (1989) : "Les grammaires d'unification", *T.A. Informations*, Paris, Klincksieck, 1989 : 1 / 2

[Numéro consacré, comme l'indique son titre, aux grammaires d'unification.]

FUCHS, C. & LE GOFFIC, P. (1992) : *Les linguistiques contemporaines ; repères théoriques*, Paris, Hachette.

[Voir le ch. $\bar{5}$: présentation de l'école transformationnelle de Z. Harris, ainsi que des travaux syntaxiques effectués par l'équipe de M. Gross, dans le cadre des "lexiques-grammaires" ; les ch. 6 & 7 : présentation des versions successives de la théorie syntaxique chomskienne, jusqu'aux développements les plus récents (version dite du "gouvernement et du liage") ; et le ch. 8 : présentation de la"grammaire lexicale-fonctionnelle" de J. Bresnan, et de la "grammaire syntagmatique généralisée" de G. Gazdar.]

GAZDAR, G. *& al.* (1985) : *Generalized phrase structure grammar*, Oxford, Blackwell.

[Exposé complet de la "grammaire syntagmatique généralisée".]

GROSS, M. (1975) : *Méthodes en syntaxe*, Paris, Hermann.

[Exposé de la méthodologie de constitution d'un "lexique-grammaire" des verbes français.]

HARRIS, Z. (1968) : *Mathematical structures of English*, New-York, Wiley ; trad.fr. (1971) : *Structures mathématiques du langage*, Paris, Dunod.

[Premier exposé d'ensemble du modèle transformationnel non-génératif de Harris.]

HARRIS, Z. (1991) : *A theory of language and information : a mathematical approach*, Oxford, Clarendon.

[Exposé le plus récent du modèle harrissien, appliqué à la grammaire anglaise.]

JOSHI, A. (1987) : Introduction to Tree Adjoining Grammar, dans A. Manaster Ramer (ed.) : *The Mathematics of Language*, Amsterdam, Benjamins.

[Exposé des "grammaires d'arbres adjoints" (TAGs).]

KAPLAN, R. & BRESNAN, J. (1982) : Lexical-functional grammar : a formal system for grammatical representation, dans J. Bresnan (ed.) : *The mental representation of grammatical relations*, Cambridge Mass., M.I.T. Press, 173-281.

[Exposé d'ensemble de la syntaxe formelle non transformationnelle de Bresnan, dite "grammaire lexicale-fonctionnelle".]

LECLERE, C. & SUBIRATS-RÜGGEBERG, C. (1991) : A bibliography of studies on lexicon-grammar, *Linguisticae Investigationes*, XV : 2, Amsterdam, Benjamins, 347-409.

[Bibliographie des principaux travaux réalisés sur diverses langues dans la perspective des lexiques-grammaires du LADL de M. Gross.]

MILLER, P. & TORRIS, T. (1990) : *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Paris, Hermès.

[Recueil d'articles présentant les principaux formalismes récents utilisés pour le traitement automatique de la syntaxe.]

MONTAGUE, R. (1970) : English as a formal language, repris dans R.H. Thomason (1974) : *Formal philosophy : selected papers of Richard Montague*, Yale University Press.

[Formalisme d'analyse de l'anglais recourant à la théorie des grammaires catégorielles.]

MOORTGAT, M. (1989) : *Categorial investigations : logical and linguistic aspects of the Lambek calculus*, Dordrecht, Foris.

[Exposé de la grammaire catégorielle généralisée. — version française adaptée dans P. Miller et T. Torris, 1990, 127-182.]

POLLARD, C. & SAG, I. (1987) : Information-based syntax and semantics, Stanford University, CSLI.

[Exposé de la théorie de la "grammaire syntagmatique dirigée par la tête" (HPSG).]

UZKOREIT, H. (1986) : Categorial unification grammars, *Proceedings of Coling 86,* Bonn,187-194.

[Exposé de la grammaire d'unification catégorielle. — version française révisée dans P. Miller et T. Torris, 1990, 183-205.]

3. Théorie (algébrique) des langages :

CHOMSKY, N. (1956) : Three models for the description of language, *IRE Transactions on information theory* ; trad. fr. (1968) : Trois modèles de description du langage, *Langages*, 9, Paris, Larousse, 51-76.

[Un des tout premiers articles cherchant à évaluer, pour une description syntaxique de l'anglais, les différents types de grammaires de phrase : chaîne de Markov à nombre fini d'états, grammaires syntagmatiques ("grammaires de constituants"), grammaires transformationnelles.]

GROSS, M. & LENTIN, A. (1970) : *Notions sur les grammaires formelles*, Paris, Gauthier-Villars.

[Bien que déjà ancien, cet ouvrage constitue la présentation de référence introduisant aux grammaires formelles et à leurs propriétés logico-algébriques.]

ROUAULT, J. (1987) : *Linguistique automatique : applications documentaires*, Berne, Lang.

[Voir le ch. 6 : "Eléments de théorie des langages".]

SELLS, P. *& al.* (eds) (1991) : *Foundational issues in natural language processing*, Bradford, Cambridge Mass. MIT Press.

[Voir en particulier le ch. 2 : "The convergence of midly context – sensitive grammar formalisms".]

4. Analyseurs syntaxiques automatiques :

ABEILLE, A. (ed.) (1991) : "Analyseurs syntaxiques du français", *T.A. Informations*, 1991 : 2, Paris, Klincksieck.

[Voir en particulier pp. 106-120 les résultats de six analyseurs du français sur un corps expérimental de 30 phrases - tests.]

BASCHUNG, K.(1992) : *Grammaires d'unification à traits et contrôle des infinitives en français*, Clermont-Ferrand, ADOSA.

[Introduction aux deux modèles de la "grammaire syntagmatique généralisée" et de la "grammaire catégorielle d'unification" ; évaluation de ces deux modèles pour un traitement automatique des infinitives en français, et présentation d'un système implémenté.]

BERWICK, R. (1991) : Principle-based parsing, dans P. Sells *& al.* (eds) : *Foundational issues in natural language processing*, Cambridge Mass., M.I.T. Press, 115-226.

[Exemple d'un analyseur de l'anglais fondé sur la théorie chomskienne du "gouvernement et du liage" (GB).]

CHANOD J.P. (1991) : Analyse automatique d'erreurs : stratégie linguistique et computationnelle, *Actes du colloque Informatique et Langue Naturelle*, LIANA, Université de Nantes, 55-71.

[Description sommaire de FROG, analyseur syntaxique du français à large couverture développé par IBM-France, destiné en particulier à la correction orthographique.]

COULOMBE, Cl. (1991) : Présentation de Machina Sapiens ; et : Les qualités attendues d'un correcteur orthographique et syntaxique, *Traitement automatique de la langue et industries de l'information (Salon international des industries de la langue, novembre 91)*, Paris, OFIL, Catalogue des exposants 24-25 ; et Actes du Colloque "Problématiques 1995" 56-57.

[Présentation succinte d'un analyseur syntaxique du français à large couverture, développé par Machina Sapiens Inc., destiné à la correction orthographique et à l'apprentissage du français (système "Exploratexte").]

FAY-VARNIER, Ch. *et al.* (1991) : Modules syntaxiques des systèmes d'analyse du français, *Technique et Science Informatiques*, Paris, Bordas, 403-425.

[Synthèse technique bien documentée des différents systèmes d'analyse syntaxique automatique développés en France.]

GARDENT, C. & BASCHUNG, K. (1993) : *Techniques d'analyse et de génération pour la langue naturelle*, Clermont-Ferrand, ADOSA.

[Initiation à l'implémentation de modèles grammaticaux en Prolog ; avec exercices et corrigés.]

GUILLEMIN-LANNE, S. (1991) : *Détection et correction d'erreurs syntaxiques par l'analyseur FROG*, IBM-France, Centre scientifique de Paris, étude F-156.

[Etat de l'art sur les correcteurs grammaticaux, suivi d'une présentation de la correction orthographique dans l'analyseur FROG d'IBM-France : exemples de correction de fautes sur les participes passés et les infinitifs en français.]

HERZOG, O. & ROLLINGER, C.R. (éd.) (1991) : *Text understanding in LILOG*, Berlin, Springer, Coll. "Lecture notes on artificial intelligence".

[Recueil d'articles concernant le système de compréhension LILOG élaboré par IBM-Allemagne. Voir pp. 33-101 pour la description de l'analyseur, et pp.183-240 pour les formalismes syntaxiques utilisés.]

KAPLAN, J. (1979) : A general syntactic processor, dans Rustin (ed.) : *Natural language processing*, New-York, Algorithmic Press.

[Analyseur recourant à la technique des "charts".]

KAY, M. (1979) : Functional grammars, *Proceedings of the 5th annual meeting of the Berkeley linguistic society*, Berkeley.

[Le premier exemple de système d'analyse implémentant le principe d'unification]

MARCUS, M. (1980) : *A theory of syntactic recognition for natural language*, Cambridge Mass., M.I.T. Press.

[Un exemple d'analyseur déterministe.]

MC.CLELLAND, J.L. & KAWAMOTO, A.H. (1986) : Mechanisms of sentence processing : assigning roles to constituents of sentences, dans J.L. Mc.Clelland & D.E. Rumelhart (eds.) : *Parallel distributed processing*, vol. 2, Cambridge Mass., M.I.T. Press, 272-325.

[Exemple d'une approche connexionniste du problème syntaxique du rattachement des groupes prépositionnels.]

SAGER, N. (1981) : *Natural language information processing ; a computer grammar of English and its applications*, Reading, Addison-Wesley.

[Analyseur de l'anglais à large couverture utilisant une grammaire en chaîne enrichie d'un composant transformationnel ; application à la recherche d'informations dans le domaine médical.]

SALKOFF, M. (1973) : *Une grammaire en chaîne du français : analyse distributionnelle*, Paris, Dunod.

SALKOFF, M. (1979) : *Analyse syntaxique du français : grammaire en chaîne*, Amsterdam, Benjamins.

[Exemple d'application au français des grammaires en chaîne.]

SEGOND, F. (1991) : Un analyseur morpho-syntaxique catégoriel pour le français ; une stratégie computationnelle pour les analyses multiples dans le cas des attachements de groupes prépositionnels, *T.A. Informations*, Paris, Klincksieck, 1991 : 1, 13-26.

[Exemple d'un analyseur du français réalisé dans la perspective des grammaires catégorielles.]

SHIEBER, S.M. *& al.* (1983) : The formalism and implementation of Patr-II, dans J. Bresnan (ed.) : *Research on interactive acquisition and use of knowledge*, SRI International Artificial Intelligence Centre, Menlo Park.

[Le système d'unification le plus connu.]

STRZALKOWSKI, T. (1992) : TTP : A fast and robust parser for natural language, *Actes du 15° Congrès international COLING-92*, Nantes, vol. II, 198-204.

[Exemple d'un analyseur "robuste".]

WEHRLI, E. (1988) : Parsing with a GB grammar, dans U. Reyle & C. Rohrer (eds.) : *Natural language parsing and linguistic theories*, Dordrecht, Reidel, 177-201.

[Présentation d'un analyseur syntaxique fondé sur la théorie chomskienne du "gouvernement et du liage" (GB).]

WERMTER, S. & LEHNERT, W. (1989) : A hybrid symbolic / connectionist model for noun phrase understanding, *Connection Science*, 1 : 3, Corfax 255-272.

[Un exemple de traitement connexionniste du problème du rattachement des groupes prépositionnels.]

WOODS, W.A. (1970) : Transition network grammars for natural language analysis, *Communications of the Association for Computing Machinery*, 13 : 10, 591-606.

[L'article original présentant les ATN.]

Sur la technique des "tableaux noirs", on pourra consulter les ouvrages suivants :

ENGELMORE, R. & MORGAN, T. (eds.) (1988) : *Blackboard systems*, Reading Mass., Addison Wesley.

ERMAN, L.D. ET LESSER, V.R. (1980) : The HEARSAY-II speech understanding system : a tutorial, *Trends in speech recognition,* Prentice Hall, 361-381.

[Présentation d'une architecture de "tableau noir".]

JAGANNATHAN *& al.* (eds.) (1989) : *Blackboard architectures and applications*, San Diego, Academic Press.

5
SÉMANTIQUE

En traitement automatique, l'analyse sémantique consiste à associer à une séquence de marqueurs linguistiques (de longueur variable) une "représentation interne" censée consigner le **sens** de cette séquence. Bien que certains systèmes (dans le cas d'univers fermés très limités) opèrent directement l'analyse sémantique sans syntaxe (ou avec un minimum de syntaxe), la plupart construisent la représentation sémantique en s'appuyant sur une analyse syntaxique effectuée préalablement ou conjointement.

Nombreux sont les systèmes qui, au niveau sémantique comme au niveau syntaxique, prennent pour unité d'analyse la **phrase**. C'est pourquoi nous ne considèrerons dans le présent chapitre que les problèmes d'analyse sémantique de la phrase ; les questions relatives aux liens inter-phrastiques seront examinées au chapitre 8 consacré à la compréhension automatique de textes.

Le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les précédents. Aussi les réalisations opérationnelles sont-elles peu nombreuses, et concernent-elles des applications très **limitées**, où l'analyse sémantique se réduit de fait à l'analyse d'un domaine parfaitement circonscrit ; par contre, on est encore loin de savoir construire en grandeur réelle des analyseurs sémantiques **généraux** qui couvriraient la totalité de la langue et seraient indépendants d'un domaine d'application particulier.

Le "sens" de la phrase que vise à consigner la représentation sémantique concerne ce que l'on peut appeler le "sens propositionnel" ; celui-ci dépend directement des formes linguistiques explicitement attestées dans la phrase (il peut être calculé à partir des informations livrées par la morphologie, la syntaxe et le lexique). A ce titre, il s'oppose à la "signification pragmatique", qui

est liée aux conditions situationnelles et contextuelles d'utilisation des formes ; celle-ci résulte notamment du calcul des références ainsi que des implicites non directement décodables à partir des formes attestées : de tels calculs, qui font intervenir simultanément des connaissances linguistiques et des connaissances d'univers, reposent très largement sur des mécanismes de nature inférentielle, comme nous le verrons aux chapitres 8 et 10.

Pour comprendre les analyseurs sémantiques, nous considèrerons successivement les phénomènes sémantiques pris en compte (§ 1.), puis les formalismes utilisés pour représenter ces phénomènes (§ 2.), et enfin les techniques informatiques de traitement mises en oeuvre dans les analyseurs (§ 3.).

1. Les phénomènes sémantiques

Les séquences linguistiques (phrases) dont l'analyseur sémantique doit décrire le sens se composent d'un certain nombre de **mots** identifiés par l'analyse morphologique, et regroupés en **structures** par l'analyse syntaxique. Ces mots et ces structures constituent autant d'**indices** pour le calcul du sens : on pourrait dire, en schématisant beaucoup, que le sens résulte de la double donnée du sens des mots et du sens des relations entre mots ; autrement dit encore, que la sémantique se dédouble en une sémantique **lexicale** et une sémantique **grammaticale**.

Nous considèrerons successivement ces deux aspects de la sémantique, tout en rappelant l'interdépendance entre les deux.

1.1. Sémantique lexicale

La sémantique lexicale est très souvent assimilée à la sémantique des "mots pleins", comme on dit, c'est-à-dire des mots qui relèvent des grandes catégories comme le substantif, le verbe ou l'adjectif (par opposition aux "mots vides", mots outils grammaticaux ou mots fonctionnels comme les prépositions, les articles, etc., que l'on a plutôt tendance à traiter comme relevant de la sémantique grammaticale).

Encore faut-il préciser ce que l'on entend par "mot plein" et par "mot vide". D'une part en effet il existe, à l'intérieur d'une catégorie de mots dits pleins (comme par exemple les verbes), des termes qui apparaissent davantage comme des "opérateurs grammaticaux" que comme des unités sémantiquement autonomes (ainsi les "verbes-supports", comme par exemple *prendre* dans *prendre courage,* ou *donner* dans *donner de l'aide à quelqu'un*). D'autre part à l'inverse il existe, parmi les catégories de mots dits vides, des termes à contenu lexical fort (adverbes formés sur des adjectifs comme *violemment*, locutions prépositionnelles comme *au milieu de*, locutions conjonctives comme *bien que*, etc.).

Aux mots lexicaux, on cherche à associer une représentation **conceptuelle** censée en décrire le sens. Deux grands types de structuration des sens lexicaux sont utilisés en traitement automatique : d'une part des **décompositions sémantiques** des unités lexicales, d'autre part l'inscription des unités lexicales au sein de **réseaux sémantiques.**

1.1.1. Décompositions sémantiques

Avant de présenter les décompositions sémantiques proprement dites, nous rappellerons que certains traitements recourent dès la syntaxe à des **traits sémantiques.**

Le recours à des traits sémantiques, déjà évoqué au chapitre précédent lors de la présentation des analyseurs syntaxiques, consiste à associer aux unités lexicales un certain nombre de traits binaires (telle unité possède ou ne possède pas tel ou tel trait, comme par exemple "humain", "animé", etc.). Comme on l'a vu (cf. *supra,* chapitre 4, § 2.2.) , cela permet de bloquer dès la syntaxe certains regroupements qui ne seraient pas pertinents du point de vue sémantique. Ainsi à la séquence *un professeur de droit allemand* (qui est réellement ambiguë, car l'adjectif *allemand* peut se rapporter aussi bien à *professeur* qu'à *droit*) s'oppose la séquence *un professeur de judo blond* (qui ne l'est pas, car l'adjectif *blond* ne peut se rapporter qu'à *professeur*) ; pour éviter de construire le regroupement *(judo blond)*, on indiquera par exemple que *blond* ne peut qualifier qu'un nom ayant les traits "+animé" "+humain", traits que l'on associera dans le dictionnaire au nom *professeur*, mais pas au nom *judo.*

Rappelons toutefois que le recours à de tels traits **binaires** est très rigide et ne permet pas de rendre compte d'un certain nombre d'emplois des termes. Ainsi, si l'on impose sur *blond* les contraintes qui viennent d'être précisées, alors des séquences comme *le sable blond* ou *les blés blonds* ne pourront pas être reconnues ; de même, si l'on indique dans le dictionnaire qu'un verbe comme *dévorer* n'admet comme objet qu'un nom ayant le trait "+comestible", alors on sera amené à rejeter comme sémantiquement mal formées des séquences pourtant parfaitement acceptables, comme *dévorer un livre* ou *dévorer quelqu'un des yeux.* Ce sont donc tous les emplois que l'on juge s'écarter d'un sens littéral de base (généralement concret et prototypique) qui se trouvent exclus par le recours à des traits sémantiques binaires. Pour éviter ces inconvénients du binarisme, certaines solutions plus souples ont été proposées, comme par exemple le recours à une "sémantique préférentielle", où les compatibilités seraient décrites de façon relative et non pas absolue (cf. Y. Wilks 1975 et 1978).

Le recours aux traits sémantiques joue donc un rôle contextuel, puisqu'il vise à décrire les compatibilités et incompatibilités sémantiques entre unités lexicales. S'il se rencontre dès le niveau syntaxique dans certains analyseurs, il se trouve également au niveau sémantique, lors de l'interprétation des relations syntaxiques (cf. *infra*, § 1.2.3.).

Passons à présent aux véritables **décompositions sémantiques**. L'idée d'une décomposition des unités lexicales en **primitives** sémantiques est simple : le sens d'une unité lexicale serait comparable à une molécule composée d'atomes, c'est-à-dire d'unités de sens élémentaires.

Ce courant s'apparente à celui que l'on connaît en linguistique sous le nom de **sémantique componentielle** (cf. les travaux de B. Pottier ou de F. Rastier), qui, adoptant en sémantique la démarche de la phonologie, recherche les "sèmes" (traits sémantiques minimaux, dont la composition constitue le sens des unités lexicales) : ainsi, dans le champ sémantique des sièges, le sème *"avoir des bras"* caractérise-t-il positivement le *fauteuil* et le *canapé*, et négativement la *chaise* et le *tabouret*, cependant que le sème *"avoir un dossier"* caractérise positivement la *chaise*, le *fauteuil* et le *canapé*, et négativement le *tabouret*, etc. Rappelons que cette approche a également connu des développements spécifiques en psycho-linguistique cognitive (cf. par exemple J-F. Le Ny 1979 ch. IV).

En traitement automatique, c'est moins l'exigence de cohérence théorique que la recherche d'une solution opératoire qui prévaut : les primitives sémantiques n'ont en général pas de définition rigoureuse, et sont souvent choisies au coup par coup, de façon assez empirique, en fonction du domaine d'application ; il s'agit moins de traits que d'opérateurs élémentaires, et ce sont le plus souvent des unités lexicales de nature prédicative qui font l'objet de telles décompositions (à cet égard, la démarche n'est pas sans rappeler celle qu'avaient adoptée, en linguistique, les tenants du courant dit de la "sémantique générative" vers la fin des années 60 : cf. la célèbre analyse de *tuer* en *"FAIRE mourir"* ; pour une présentation synthétique de ce courant linguistique, nous renvoyons à M. Galmiche 1975).

Il faut ici distinguer deux approches très différentes, tant dans leurs objectifs que dans leurs présupposés théoriques.

La première tente de décrire toutes les unités lexicales de la langue (ou plus modestement d'un domaine sémantique spécifique) à l'aide d'un **ensemble minimal** de primitives, qui formeraient ainsi les unités sémantiques de base dont tout terme lexical serait une combinaison. Cette "version forte" de la décomposition sémantique se heurte à de nombreuses difficultés. D'abord dans le choix de ces primitives : toutes les tentatives "généralistes" de construire un tel système (et elles ne datent pas d'aujourd'hui : on peut remonter au moins à Leibniz !) se sont avérées irréalistes. C'est bien sûr plus abordable dans un domaine strictement limité ; mais quel est alors l'intérêt pour le traitement de la langue : J. Pitrat (1985, p.14), par exemple, a montré que même dans des petits textes traitant *a priori* d'un thème très étroit (les commentaires techniques de parties d'échecs) la richesse de la langue utilisée rendait vaine toute approche réductionniste. Un autre problème majeur provient de la **polysémie**, phénomène massif, en particulier pour les unités lexicales les plus fréquentes. Le seul traitement possible dans ce cadre est de

représenter les différentes acceptions d'un mot comme autant d'unités distinctes, ce qui présente le double inconvénient de masquer l'unité de sens inhérente à la polysémie, et d'augmenter sans fin le nombre d'entrées lexicales, provoquant ainsi une prolifération d'ambiguïtés artificielles.

La deuxième approche est moins ambitieuse et plus raisonnable : il s'agit de définir pour chaque mot un ensemble de traits qui le caractérise, en particulier en le distinguant des mots de sens voisin. Le nombre de ces traits n'est **pas limité** : on ne cherche pas à proprement parler à en faire des primitives. Dans ce cadre, la polysémie peut être prise en compte plus facilement : il suffit de considérer que tous les traits afférents à un polysème ne sont pas forcément présents dans tous les emplois du mot. Cela implique alors un calcul qui détermine, en fonction du contexte du mot dans la phrase, quels sont les traits "activés" et les traits "effacés", et donc quel sens précis prend le polysème dans la phrase (cf. par exemple F. Rastier 1987).

1.1.2. Réseaux sémantiques

Ce type de représentation a été à l'origine élaboré dans le domaine de la psychologie ; il s'agissait de rendre compte de la façon dont les sujets catégorisent et mémorisent les concepts : cf. R. Quillian (1968). Très vite, les réseaux sémantiques ont connu une grande vogue dans les milieux de l'intelligence artificielle. L'idée ici n'est plus de décomposer une unité lexicale en une structure sémantique plus petite, mais d'inscrire l'unité lexicale dans une structure sémantique plus vaste, de telle sorte que le sens de l'unité résulte de la place qu'elle occupe dans la structure, et des relations qu'elle entretient avec les autres unités de cette structure. Formellement, les réseaux sémantiques sont des **graphes**, formés de **noeuds**, qui représentent les unités et sont reliés par des **arcs**, qui représentent les relations entre unités.

La relation la plus évidente à considérer est la relation hiérarchique entre hyperonymes et hyponymes : un *merle* est un *oiseau*, un *oiseau* est un *animal*, etc., créant ainsi une structure **arborescente**. Les relations de synonymie et d'antonymie sont aussi envisageables ; en outre, une relation de "quasi-synonymie", non transitive (A et B peuvent être quasi-synonymes, B et C aussi, sans que A et C le soient) permet d'établir une notion topologique de proximité sémantique, qui semble intuitivement correspondre à une réalité de la langue.

Mais cela ne saurait suffire. D'une part, il existe peu de domaines descriptibles par une véritable taxinomie, comme celle des êtres vivants, et les essais (nombreux : l'idée remonte cette fois-ci à Aristote) de structurer ainsi un lexique général se sont avérés vains. D'autre part, de toutes façons, ces relations ne peuvent à elles seules définir les mots. On a donc été amené à définir d'autres types de **relations** : en plus de liens du type "sorte-de" (qui décrit la relation d'inclusion de classes que nous venons de voir), on trouve des liens "partie-de" (une *roue* est une partie d'une *voiture*), "sert-à" (une *scie* sert à *couper*), "conséquence-de" (*être repu* est une conséquence de

manger), etc. Ainsi peut-on aboutir à une caractérisation aussi précise que l'on veut d'une unité du réseau.

Les difficultés restent cependant nombreuses. La définition précise d'un lien ne va pas de soi et peut recouvrir des relations fort différentes (une *chambre* n'est pas une partie d'un *appartement* de la même manière qu'un *lit* serait une partie d'une *chambre*, ou que la *foudre* serait une partie d'un *orage*, ou encore qu'un *instituteur* serait une partie de l'*Education Nationale* !). A l'inverse, l'inflation du nombre de types de liens (qui sont souvent là encore choisis de façon empirique, sans base théorique rigoureuse) constitue un danger évident pour la maîtrise et la cohérence du système. Mais le problème décisif concerne la nature même des unités que l'on relie ainsi : le plus souvent, sans que cela soit toujours explicite, ce ne sont plus les termes lexicaux (ou du moins leur signifié), mais des **concepts** définis par le réseau de relations lui-même (cf. la critique de F. Rastier 1991 ch. 4 et 5). Le problème se pose alors de la mise en correspondance des mots de la langue avec ces unités conceptuelles. On retombe ainsi sur le problème de la polysémie : un même mot peut correspondre suivant ses emplois à un grand nombre de noeuds du réseau, et il faut se donner les moyens de calculer, en fonction de la phrase dans laquelle le mot apparaît, à quelle unité conceptuelle on a affaire.

Signalons enfin que décompositions sémantiques et réseaux sémantiques ne sont pas forcément contradictoires. En effet si l'on suppose que l'on a réalisé une décomposition en traits, on peut en déduire automatiquement un réseau correspondant : si une unité A possède tous les traits d'une unité B, plus quelques traits spécifiques, c'est que B est un hyperonyme de A. De même le pourcentage de traits communs à deux unités peut servir à définir une proximité sémantique dans un réseau. Réciproquement, une arborescence taxinomique peut facilement servir à définir des traits : à chaque niveau de hiérarchie, un trait supplémentaire suffit à distinguer le noeud-père des noeuds-fils, et une valeur différente de ce trait pour chaque fils permet de les différencier entre eux.

1.2. Sémantique grammaticale

Une analyse sémantique qui se réduirait à une sémantique lexicale serait, par définition, insuffisante ; il n'est, pour s'en convaincre, qu'à évoquer les lacunes et les risques d'erreur des premiers systèmes de compréhension qui ne reposaient que sur l'identification de mots clés : la compréhension des **relations** entre les mots est tout aussi importante, du point de vue sémantique, que la compréhension des mots eux-mêmes.

Les approches formelles de la sémantique grammaticale tentent de décrire le sens de la phrase en traduisant la structure syntaxique de cette phrase en une **formule logico-sémantique**. Elles disposent pour ce faire de deux types d'informations : les relations syntaxiques et la présence des marqueurs grammaticaux.

1.2.1. L'interprétation des relations syntaxiques

Elle n'est pas simple, dans la mesure où il n'y a pas de correspondance immédiate et bi-univoque entre syntaxe et sémantique. Une même structure syntaxique peut donner lieu à des représentations sémantiques différentes ; ainsi par exemple la structure "sujet verbe objet direct" renvoie-t-elle à des types de procès différents avec des rôles sémantiques différents sur les actants dans *Jean habite Paris* (état sans agentivité du sujet), dans *Jean regarde le ballon* (activité d'un sujet agent sans construction d'un état résultant sur l'objet) et dans *Jean attrape le ballon* (changement d'état de l'objet sous l'action du sujet agent). A l'inverse, plusieurs structures différentes peuvent donner lieu à une même représentation sémantique ; ainsi *Cette clé ouvre la porte ; La porte s'ouvre avec cette clé ; On ouvre la porte avec cette clé.*

Nous présenterons deux versions de ces approches : une version faible (les représentations en termes de prédicat-arguments) et une version forte (les grammaires de cas).

Les représentations en termes de **prédicat-arguments** consistent à représenter chaque proposition en termes d'une relation prédicative constituée d'un prédicat accompagné de ses arguments et d'éventuels modifieurs. Le propre de ces représentations est de simplement **numéroter** les arguments du prédicat, sans chercher à leur associer de dénomination sémantique. C'est la solution souvent adoptée en traduction automatique (cf. *infra*, chapitre 7).

Des règles d'analyse sémantique permettent, après consultation du dictionnaire, de transformer l'arborescence syntaxique en une représentation prédicative dont les noeuds sont remplis par les unités lexicales correspondant au prédicat, à ses arguments et aux éventuels modifieurs. Prenons l'exemple des trois phrases suivantes : *Marie marche* ; *Marie plait à Jean* ; *Marie donne un livre à Jean.* Ces phrases seront représentées respectivement par les trois arbres donnés dans la Figure I.

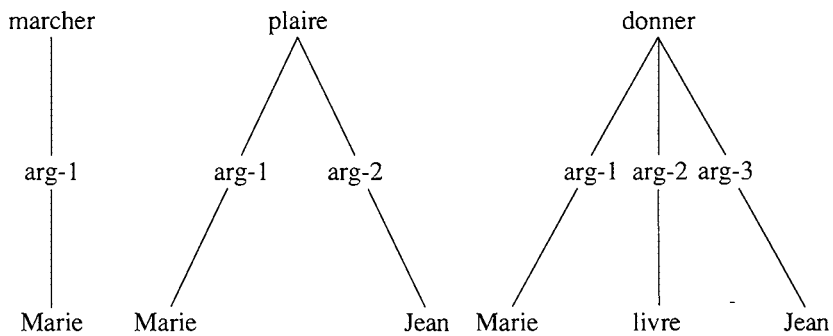


FIGURE I

Les mots grammaticaux ne figurent pas dans les noeuds : ou bien ils disparaissent de la structure sémantique (c'est le cas des prépositions gouvernées par le verbe), ou bien ils sont intégrés sous forme de valeurs de variables (ou attributs) accrochées aux noeuds (c'est le cas des temps, modes, voix et statuts assertifs accrochés au noeud prédicatif, ainsi que des déterminants accrochés aux noeuds d'arguments).

Ce type de représentation permet de ramener à une structure sémantique comparable des structures syntaxiques différentes : les diverses constructions actives (*Jean a ouvert la porte ; Jean a ouvert la porte avec cette clé ; On a ouvert la porte avec cette clé ; Cette clé a ouvert la porte*), et passives (*La porte a été ouverte; La porte a été ouverte avec cette clé*) d'un verbe comme *ouvrir* seront ramenées à une structure prédicative constituée du prédicat *ouvrir* et de trois places d'arguments qui, selon les cas, seront remplies ou non : *Jean* en argument-1, *porte* en argument-2 et *clé* en argument-3.

Dans le **dictionnaire**, l'entrée lexicale *ouvrir* comportera d'une part des informations syntaxiques concernant les différentes constructions du verbe, d'autre part des informations sémantiques concernant le nombre d'arguments du prédicat. La caractérisation du nombre d'arguments d'un prédicat, et leur identification constituent évidemment des questions délicates, auxquelles il n'existe que des réponses relativement empiriques, pas toujours très satisfaisantes du point de vue théorique (il n'existe pas de méthode suffisamment cohérente pour être totalement reproductible sans risque de variation d'une personne à l'autre).

Concernant le **nombre** d'arguments, le critère de distinction entre argument et modifieur classiquement invoqué est celui du caractère nécessaire et unique (sauf cas de coordination) des arguments, face au caractère facultatif et cumulable des modifieurs. Or si l'on accepte, comme nous l'avons fait dans l'exemple ci-dessus, que *avec cette clé* soit un argument de *ouvrir*, on sera sans doute amené de proche en proche à considérer que l'on a affaire aussi à une structure à trois arguments dans les exemples suivants :

Jean coupe sa viande avec son couteau
Jean mange sa viande avec son couteau / avec les doigts
Jean va à Paris avec sa voiture / en train / par la route / par l'autoroute.

Dès lors, combien d'arguments faut-il compter dans les exemples suivants :

Jean donne un livre à Paul pour Marie
Jean envoie un livre par la poste pour Marie
Jean envoie un livre à Paul pour Marie par le premier train ?

Comme on le voit, les critères que nous avons donnés ne sont pas faciles à mettre en oeuvre, et dans la pratique les approches oscillent entre une conception très restrictive dans laquelle tout ou presque est modifieur (ce qui enlève beaucoup de l'intérêt de la structure argumentale) et une conception

plus laxiste qui conduit à multiplier démesurément le nombre de constructions de chaque verbe.

Concernant la **numérotation** des arguments, le premier critère donné consiste à prendre comme construction de base la construction maximale du verbe au présent actif, en évitant les pronominalisations, et à numéroter les arguments dans l'ordre linéaire d'apparition. Lorsque ce critère est insuffisant, laissant le choix entre plusieurs solutions, on peut recourir, selon les cas, à une hiérarchisation *a priori* des prépositions gouvernées par le verbe (ainsi pour le verbe *traduire* on choisira comme construction de base *traduire du français en anglais* plutôt que *traduire en anglais à partir du français*) ou bien à des traits sémantiques imposés sur les arguments (ainsi on choisira comme construction de base *Jean ouvre la porte* plutôt que *La clé ouvre la porte*, en imposant le trait "+ animé" à l'argument-1 de *ouvrir*).

Là encore, ces critères ne résolvent pas toutes les difficultés. Pour nous en tenir à notre exemple, il existe en fait une gradation entre les constructions *Jean ouvre la porte* et *La clé ouvre la porte*, qui rend quelque peu arbitraire le choix dichotomique entre argument-1 et argument-2 :

Le vent ouvre la porte
Un coup de vent ouvre la porte
Une forte poussée ouvre la porte
Un coup d'épaule ouvre la porte
Un coup de barre à mine ouvre la porte
Une charge d'explosif ouvre la porte
Un pied de biche ouvre la porte.

Le trait "±animé" paraît bien maigre pour rendre compte de ces nuances !!

Par ailleurs, comme dans toute interprétation des relations syntaxiques en termes de structure prédicative, il faut préciser les cas où le verbe conjugué n'est pas l'élément qui se retrouve en place de **prédicat**. C'est ce qui se passe notamment en présence de modalités (le prédicat est le verbe à l'infinitif qui suit *pouvoir, devoir, falloir*, etc.) ou d'auxiliaires de temps (le prédicat est le verbe au participe passé qui suit *être* ou *avoir*), des verbes-supports (dans *prendre peur, avoir faim*, le prédicat est le substantif qui suit le verbe-support), des adjectifs attributs (dans *être beau*, le prédicat est *beau*), etc. A cet égard, la question de l'interprétation sémantique des verbes nominalisés est particulièrement délicate : la nominalisation fonctionne comme prédicat dans *La construction de la maison par le maçon a été laborieuse*, mais comme argument dans *Cette construction me plaît par ses proportions harmonieuses*.

Passons à présent aux **grammaires de cas**. Il s'agit pour ce type d'analyse sémantique non seulement de retrouver la relation prédicative sous-jacente à la construction du verbe, mais d'étiqueter la valeur sémantique des divers éléments (actants et circonstants) qui entourent le prédicat.

C'est un linguiste américain, Ch. Fillmore, qui, en 1968, lança ce type d'approche, en écrivant un article intitulé "The case for case". Bien qu'il soit lui-même revenu sur le contenu de cet article, et que celui-ci ait été critiqué par d'autres linguistes (lui reprochant d'avoir simplement énuméré une liste de cas de manière empirique au lieu de les avoir construits théoriquement), il n'en reste pas moins que la notion de **cas sémantique** qu'il y introduisait a connu un grand succès en traitement automatique de la sémantique grammaticale.

Pour reprendre un exemple présenté plus haut, dans *Jean ouvre la porte avec cette clé*, on dira que *Jean* est "agent", *porte* "objet" et *clé* "instrument". De même on dira que le verbe *casser* accepte l'ossature casuelle suivante: un "objet" (obligatoire), un "instrument" (facultatif) et un "agent" (facultatif). Quand un verbe possède plusieurs sens, il est décrit en termes de plusieurs ossatures casuelles ; ainsi par exemple *voler* (comme l'oiseau) acceptera un "agent", *voler* (comme l'avion) acceptera un "objet", et *voler* (au sens de *dérober*) acceptera un agent et un objet (plus éventuellement un "bénéficiaire" — encore que cette appellation ne soit pas des plus heureuses dans cet exemple !).

L'idée est qu'un petit nombre de cas sémantiques permettrait de rendre compte de toutes les constructions, et que l'on pourrait établir des **correspondances** entre les fonctions syntaxiques et les cas sémantiques, selon le verbe considéré. En pratique, il s'est avéré extrêmement difficile de mettre en oeuvre de façon opératoire une théorie des cas sémantiques, dès lors que l'on s'éloigne de petits schémas de phrases bien simples comme ceux donnés ci-dessus. Néanmoins de nombreux modèles de traitement automatique s'en sont inspirés et ont tenté de décrire les verbes en termes des cas sémantiques qu'ils acceptent ; la sélection et la caractérisation des cas reste très empirique et peu opératoire (selon les auteurs, le nombre de cas retenus varie de six à plus d'une vingtaine, la définition en est peu rigoureuse : cf. G. Sabah 1988 ch. 3).

Notons pour mémoire que bien d'autres théories des cas ont été proposées en linguistique: cf., entre autres, B. Pottier (1992 ch.X, pp.122-128).

1.2.2. Sémantique des marqueurs grammaticaux

Si, ainsi que nous venons de le voir, la structure syntaxique de la proposition se laisse (plus ou moins bien) interpréter en termes de relation prédicative, en revanche les marqueurs grammaticaux, véritables révélateurs du mode de fonctionnement sémantique de la langue, se laissent plus difficilement interpréter dans les cadres logiques retenus par les traitements automatiques.

Certains marqueurs grammaticaux spécifient des relations **intrapropositionnelles** : c'est le cas par exemple des prépositions introduisant des actants ou des circonstants du prédicat. En fait, et c'est là le problème, une même pré-

position peut jouer des rôles différents par rapport à la relation prédicative, et prendre des contenus sémantiques très variés suivant les énoncés. Prenons l'exemple de *dans* : il peut être le support de la prédication (*L'Office du Tourisme est dans le château ; Notre projet est dans les choux*), il peut introduire un argument obligatoire (*Il entra dans la pièce ; Laissez cet endroit dans l'état où vous l'avez trouvé*), et il peut spécifier un circonstanciel (*dans la maison ; dans la journée ; dans trois jours ; dans ce cas ; dans cette théorie ; ...*), en indiquant des relations spatiales, temporelles, notionnelles, etc. Encore une fois, nous retrouvons le phénomène de la **polysémie**. D'une manière générale, les marqueurs grammaticaux présentent un degré de polysémie très élevé, cependant que l'existence d'un "noyau de sens" unique est perceptible dans tous les emplois de chacun d'eux. C'est ce qui conduit un certain nombre de théories linguistiques à donner une place centrale à la représentation de ces noyaux de sens et à l'étude du fonctionnement de ces unités dans les énoncés. Dans les grammaires cognitives américaines (G. Lakoff 1987, R. Langacker 1987 / 1991, L. Talmy 1988), le noyau de sens est représenté par un schéma iconique, dans la théorie de l'énonciation de A. Culioli (1990), il correspond à une opération, etc.

D'autres marqueurs grammaticaux spécifient des relations **interpropositionnelles** (par exemple des conjonctions de subordination). Leur interprétation dans un cadre logique n'est pas aisée ; en particulier les ramener aux opérateurs de la logique des propositions apparaît comme extrêmement réducteur. Pour ne prendre qu'un exemple, le connecteur *si* ne correspond à l'opérateur logique d'implication "si p alors q" que dans l'un de ses emplois, celui que l'on a dans *Si tu viens, nous irons nous promener*, mais il ne lui correspond pas dans des emplois comme : *Si le public a applaudi, c'est que la pièce était bonne ; S'il me rencontrait, il me saluait toujours ; Si je comprends tes raisons, je ne les approuve pas ; Si tu as soif, il y a de la bière dans le frigo.*

Il existe encore d'autres types de marqueurs grammaticaux : ce sont les "**indices référentiels**" qui, comme les déterminants, les temps, les modalités, les personnes, etc., assignent aux différents constituants d'une proposition certaines valeurs (valeurs temporelles, modales, etc.), leur permettant ainsi de **référer** au monde, et donnant à la proposition le statut non pas seulement de phrase prédicativement bien formée, mais d'**énoncé** signifiant (nous les réévoquerons au chapitre 8, § 1.2.). Ces indices référentiels constituent, on le sait, l'objet d'étude privilégié des linguistiques de l'**énonciation.** Dans les traitements automatiques en revanche, la distinction entre "valeur référentielle" (relevant d'une sémantique du système de la langue) et "référence effective" (relevant de la pragmatique) est très largement ignorée.

Si le calcul de la référence effective relève de la pragmatique (comme par exemple de savoir de quel chien particulier "Médor" ou "Fido" je parle en disant *ce chien*), en revanche le calcul des valeurs linguistiques qui permettent

cette assignation de référence relève, lui, de la sémantique (ici le fait de savoir qu'en français le démonstratif *ce* permet de référer à une occurrence particulière soit par ostension directe : "ce chien que je montre du doigt, qui est dans notre champ de vision", soit par anaphore renvoyant non plus à la situation mais au contexte : "ce chien dont je viens de parler"). Pour que le mot *chien* prenne une référence effective, il faut en effet parler de *ce chien*, de *mon chien*, de *le chien du voisin*, etc. De même pour qu'un prédicat renvoie à un événement du monde, il faut l'ancrer dans le temps (cela se passe au moment ou je parle, cela s'est passé avant, se passera plus tard), lui donner une valeur aspectuelle (événement en train de se dérouler au moment considéré, ou achevé), et aussi une valeur modale (c'est vrai ou faux, ou bien c'est une hypothèse, ou encore une question, c'est probable, éventuel, certain, etc.). Longtemps ignorées des traitements automatiques, ces questions font actuellement l'objet de tentatives de formalisations, en particulier dans des cadres logiques.

La difficulté est énorme, comme quelques exemples vont le montrer. Prenons d'abord la **détermination**. Les articles *un* et *le* peuvent prendre une valeur générique (*Le merle est un oiseau de la famille des passereaux* ; *Un merle a le plumage plus foncé qu'une grive*) et ils correspondent alors, en partie tout au moins, au quantificateur "universel" de la logique ("pour tout x"). Mais ils peuvent aussi prendre une valeur spécifique (*C'est un merle qui niche dans notre jardin* ; *Le merle de notre jardin siffle à longueur de journée*), et dans ce cas ils seraient plus proches du quantificateur "existentiel" ("il existe un x"). De plus, dans la plupart de leurs emplois, *le* et *un* ne sont pas interchangeables (même pour la valeur générique ; cf. G. Kleiber (1990): dans *Le castor a été introduit par les autonomistes en Alsace en 1925*, l'article *le* ne peut pas être remplacé par *un*). Ainsi ces deux marqueurs sont incontestablement différents, mais leur différence n'a rien à voir avec le type de distinction que l'on opère en logique.

Ce type d'inadéquation se retrouve aussi dans le domaine des valeurs **modales**. Ainsi les verbes modaux *pouvoir* et *devoir* peuvent tous les deux exprimer des valeurs épistémiques assez proches de l'éventualité et de la probabilité (*Il peut pleuvoir demain* ; *Il doit faire beau sur la côte*) qui se traduiraient, comme on le verra plus bas, dans certaines logiques "modales" (cf. *infra*, § 2.2) par un même opérateur de "contingence" ("il est possible que"). En revanche, dans certains de leurs emplois, ils prennent des valeurs déontiques très contrastées (*Tu dois te taire* ; *Tu peux fumer*) et correspondent alors, dans d'autres logiques modales, à des opérateurs distincts ("il est obligatoire de" et "il est autorisé de"). D'autres valeurs (comme la capacité pour *pouvoir*: *Le pingouin peut voler, contrairement au manchot*) s'interprètent dans un cadre logique comme un simple prédicat (une propriété) sans avoir besoin de faire appel à une logique modale.

Le domaine de la **temporalité** fournit aussi un bon exemple. Il existe des logiques dites "temporelles", qui permettent d'exprimer une notion de valeur

de vérité dépendante du temps (à l'aide d'opérateurs tels que "il a été vrai que" et "il sera toujours vrai que" : cf. *infra* § 2.2.). Mais en fait les temps grammaticaux, malgré leurs noms, ne peuvent pas être mis en correspondance biunivoque avec les concepts de présent, passé, futur : les marques de présent ou de futur peuvent référer à des événements passés (*Evariste Galois naît en 1811, et il mourra à 21 ans dans un duel*) et inversement (*Je pars mardi, après avoir assisté à la réunion le matin*). En outre, dans bien des emplois, les temps verbaux marquent autre chose que le temps (*Tout homme est mortel ; Si tu savais... ; Il se sera trompé de jour*). Cela ne siginifie pas bien sûr que l'utilisation des temps grammaticaux relève de l'aléatoire des locuteurs ou de l'arbitraire des grammairiens. Au contraire, elle obéit à une logique interne à la langue, qui permet d'expliquer la pluralité des valeurs référentielles obtenues : mais comme le montrent les innombrables travaux linguistiques consacrés à ces questions, les régularités observées relèvent de mécanismes complexes (faisant appel à des notions de points de vue et de points de référence multiples, de dynamique des procès, de changement de repères, etc.), qui ne sont pas facilement modélisables dans des cadres logiques — lesquels, malgré tous les efforts en ce sens depuis plusieurs années, restent encore trop réducteurs.

La **polysémie** des marqueurs grammaticaux, dont l'omniprésence a été soulignée tout au long de ce survol, conduit souvent à des **ambiguïtés** proprement sémantiques (non réductibles à des ambiguïtés de construction syntaxique). Elles sont "globales" au sens où elles ne peuvent pas être levées dans le cadre de l'énoncé isolé : seul un contexte plus large peut permettre la compréhension. Sans chercher le moins du monde à être exhaustif, citons-en quelques-unes :

- détermination : la relative peut être restrictive ou descriptive dans :

  Les élèves qui ont chahuté seront punis (= "Les élèves, parce qu'ils ont..." ?
  ou "Ceux des élèves qui ont.... " ?)

- quantification : on ne peut pas trancher entre unicité et distributivité dans :

  Tous les élèves ont traité une question du problème (tous la même ?)

- aspect : *encore* est-il répétitif ou duratif dans :

  Jean est encore malade (= "toujours pas guéri" ? ou "de nouveau malade" ?)

- modalité : l'imparfait peut décrire un passé réel ou un passé fictif dans :

  Avec ce remède de cheval, Jean était rétabli en un clin d'oeil (= "aurait été" ? ou "a été" ?)

- référence : on peut avoir "opacité référentielle" ou non dans :

  Jean veut boire cette liqueur empoisonnée (Jean veut-il boire une liqueur, que je sais, moi, être du poison ou Jean a-t-il envie de se suicider ?).

2. Les formalismes de représentation

Les deux grandes familles de formalismes utilisés pour construire les représentations sémantiques sont d'une part les structures attributs-valeurs, et d'autre part les formalismes logiques (dont la diversité est par ailleurs très grande) ; par-delà ces voies classiques, certaines approches alternatives se font jour actuellement.

2.1. Les structures attributs-valeurs

La structure de base la plus utilisée dans les représentations sémantiques est une structure de données très simple : chaque unité est composée d'une liste de **couples attribut-valeur.** Nous avons déjà rencontré cette structure dans certains formalismes syntaxiques (les structures de traits dans les grammaires d'unification — cf. *supra*, chapitre 4 § 4.4.). Dans une telle liste, la valeur associée à un attribut peut être simple (binaire, quantitative, qualitative) ou complexe : une autre unité est alors reliée à la première par l'intermédiaire de cet attribut. On peut représenter ce système par un **graphe**: chaque unité est un noeud du graphe, chacun de ses attributs un arc issu de ce noeud, les valeurs simples représentant des noeuds terminaux (aucun arc ne part de ces noeuds). Par exemple, supposons que l'on ait un noeud *merle* qui contienne dans sa liste d'attributs-valeurs les couples (sorte-de, *oiseau*) (couleur, *noir*) (propriété, *siffler*) où *noir* est une valeur qualitative simple, tandis que *oiseau* et *siffler* sont d'autres unités possédant elles-mêmes des attributs sorte-de, propriété, ... La petite portion de système que nous venons de décrire peut donc se représenter de deux façons, de manière totalement équivalente (cf. Figures II et III).

On voit bien sur cet exemple l'intérêt essentiel de ces structures : elles peuvent tout aussi bien représenter des **décompositions** sémantiques (les attributs sont alors des traits) que des **réseaux** sémantiques (les attributs sont alors des relations entre unités). Mieux, ces deux types de représentation y sont naturellement mêlés. Un certain nombre de mécanismes contribuent à faciliter le passage de l'un à l'autre. Le plus important est l'**héritage**, défini le long de liens hiérarchiques comme "sorte-de", qui permet d'ajouter "virtuellement" certains couples attribut-valeur d'une unité à une autre. Ainsi, *merle* pourra hériter la propriété de *voler* parce que c'est une sorte d'*oiseau*. Ce mécanisme peut être tempéré par une gestion des exceptions : si l'on définit une unité *autruche* qui soit aussi une sorte d'*oiseau*, on peut, en l'indiquant explicitement, éviter que *autruche* n'hérite de la propriété de *voler*. C'est la base de ce que l'on appelle les valeurs **par défaut** : elles ne sont utilisées qu'en l'absence d'une information explicite qui les contredise. L'héritage par défaut permet donc une grande économie dans l'acquisition et la gestion des informations. Il est aussi la base dans certains systèmes (par des mécanismes divers) de la représentation d'une notion de **prototype** et de valeur typique.

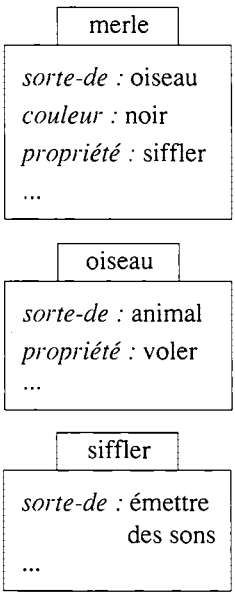


FIGURE II

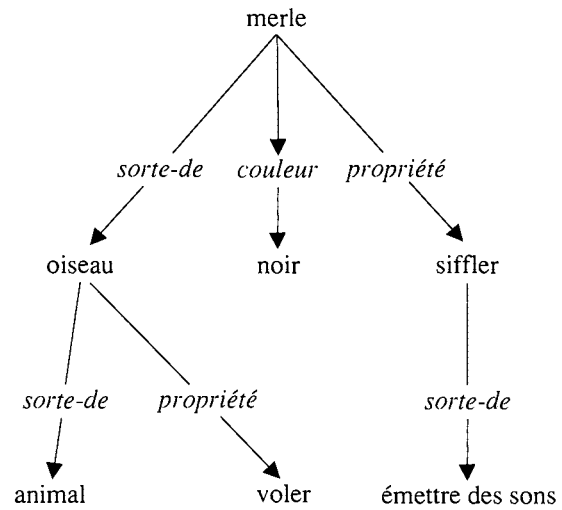


FIGURE III

Le principe que nous venons d'exposer est très général. En fait beaucoup de formalismes, très différents dans leur conception et dans leur puissance d'expression, ont été développés sur cette base. Les premiers sont nés à la fin des années 70 : les réseaux de propagation de marqueurs de S. Fahlman (1979), les réseaux partitionnés de G. Hendrix (1979), le système KL-ONE de R. Brachman & J. Schmolze (1985), etc. (pour une description rapide, voir G. Sabah 1988 ch. 7). Depuis, une foule de systèmes ont été conçus, et la généralisation des langages à objets a en quelque sorte banalisé le principe de base sous-jacent. On trouvera dans J. Sowa (ed., 1991) un exposé des orientations actuelles de la recherche dans ce domaine.

Ces systèmes sont en fait des systèmes généraux de **représentation de connaissances**, et comme nous l'avons dit, il règne une certaine confusion sur le statut exact des unités : parfois ce sont les unités lexicales elles-mêmes dont le système cherche à capter le signifié, mais plus souvent ce sont des concepts auxquels sont attachées toutes sortes de connaissances. Ce qui est problématique ici, c'est la distinction entre connaissances lexicales et connaissances encyclopédiques, sur laquelle nous aurons l'occasion de revenir (cf. *infra*, chapitre 8 § 2.1.).

Un certain nombre de systèmes utilisent les **graphes**, non seulement pour représenter les connaissances lexicales (leur "dictionnaire", en somme), mais aussi pour la représentation sémantique des **phrases** elles-mêmes. Le graphe

sert donc de format de sortie à un analyseur à qui l'on donne en entrée une phrase. Un des premiers systèmes de ce type a été élaboré par R. Schank (1972). Celui-ci a opté pour l'utilisation d'un système de primitives, qu'il a appelé "primitives de **dépendance conceptuelle**". Pour R. Schank, toutes les actions sont décomposables à l'aide de onze concepts de base, de type "appliquer une force à quelque chose", "attraper un objet", "introduire quelque chose à l'intérieur de quelque chose", "changer la position d'un objet", "produire un son", "transférer des informations d'un individu à un autre", etc. A ces actions sont associés des attributs, dont certains, en nombre très limité, correspondent aux cas d'une grammaire de cas, et d'autres à des modalités (temps, statut assertif, etc. — là encore, en nombre très limité : le système est très frustre) ; de même les objets sont définis par des listes attributs-valeurs. Une phrase est alors représentée par une combinaison de ces primitives. Par exemple, la Figure IV illustre le graphe correspondant à la phrase *Pierre dit à Paul que Jean a donné un livre à Marie.* L'intérêt de ce système est avant tout d'avoir été l'une des seules tentatives d'implémenter la "version forte" de la décomposition sémantique (cf. *supra*, § 1.1.1.).

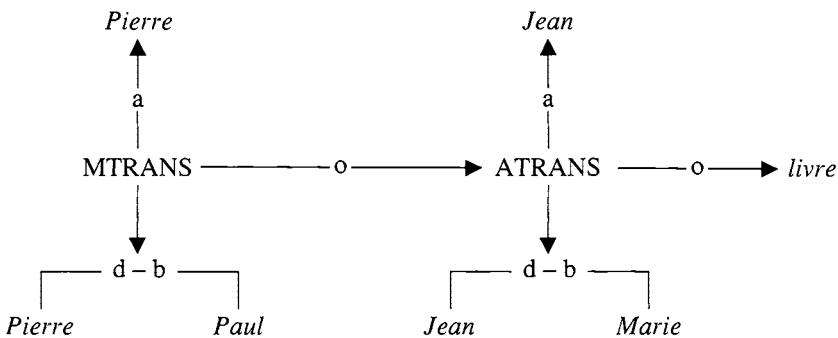


FIGURE IV

(sur la figure *ATRANS* dénote la primitive de transfert de propriété et *MTRANS* la primitive de transfert d'information, *a* désigne le rôle d'acteur, *o* celui d'objet, et *d-b* le couple de rôles donateur-bénéficiaire).

Il faudrait citer aussi, dans ce courant théorique, la "sémantique préférentielle" de Y. Wilks (1975 et 1978) basée sur une centaine de primitives, chaque mot étant décrit par une formule combinant ces primitives, les phrases étant à leur tour des combinaisons de formules.

Dans un tout autre cadre théorique, d'autres systèmes utilisent des réseaux sémantiques "classiques" pour représenter dans un même formalisme connaissances lexicales et sens des phrases. Le point de départ, dans beaucoup de systèmes de ce type, est de distinguer des noeuds qui représentent des **classes** et des noeuds qui représentent des **individus** spécifiés, liés à une classe par un

lien d'appartenance ou d'instanciation (là encore les langages à objets ont banalisé cette notion). A une phrase comme *Jean ira à Paris demain* est alors associé un sous-graphe dans lequel un noeud représente l'événement lui-même, comme instance de la classe des voyages, relié d'une part au noeud représentant *Jean*, de la classe des humains, et au noeud *Paris*, de la classe des villes. La date de l'événement pourra être, suivant les approches, soit une simple valeur d'un attribut particulier aux événements, soit un noeud d'une classe d'entités temporelles. La plupart des formalismes de réseaux sémantiques permettent d'implémenter cette idée sous une forme ou sous une autre.

On peut aussi sans doute classer dans cette catégorie les **graphes conceptuels** de J. Sowa (1984), qui connaissent aujourd'hui une popularité certaine, encore que la notion de type utilisée dans ce système soit plus proche de celle des logiques typées (voir *infra*, § 2.2.) que de celle de classe. L'effort de J. Sowa a surtout porté sur la mise au point d'un formalisme de réseaux dont la puissance d'expression soit au moins aussi forte que celle de la logique classique. Pour cela, il introduit une distinction entre concepts "génériques" et concepts "individuels" de même type. Sans entrer dans le détail ici, on peut dire très grossièrement que les concepts génériques jouent un rôle analogue à celui des variables en logique et que des opérations sur les graphes "correspondent" à des déductions.

2.2. Les formalismes logiques

C'est de la logique que provient l'autre grande famille de formalismes qui ont été proposés pour représenter les phénomènes sémantiques. Il ne saurait être question de présenter dans le cadre de cet ouvrage les diverses théories logiques (chacune d'elles mérite un ouvrage entier), ni d'exposer les relations entre logique(s) et linguistique(s) (c'est aussi une très longue histoire). Nous allons donc nous contenter, en restant à un niveau "intuitif", avec tout ce que cela comporte de raccourcis et de déformations, de donner une idée de la manière dont ces théories peuvent servir d'outils de représentation dans le cadre du traitement automatique.

Voyons d'abord comment des formules logiques peuvent représenter des connaissances **lexicales**. L'idée de base est de faire correspondre à chaque signifié (ou à chaque concept : la confusion que nous avons évoquée plus haut reste tout aussi grande) un prédicat possédant un nombre fixe d'arguments. Ainsi, pour simplifier à l'extrême, à un nom comme *merle* est associé le prédicat unaire MERLE(x), et à un verbe comme *donner* le prédicat ternaire DONNER(x,y,z). Ces prédicats ne sont pas des définitions des concepts correspondants. Ils doivent simplement être entendus de la manière suivante : étant donné un monde dans lequel existent des individus spécifiés A, B, C..., on peut faire correspondre (créant ainsi un "modèle") au prédicat MERLE une propriété dans ce monde dont on peut dire si elle est vraie ou fausse pour A, pour B, pour C, etc. ; de même

pour DONNER, en considérant cette fois tous les triplets possibles d'individus. De tels prédicats ne peuvent donc servir à représenter des connaissances lexicales que si on les relie par l'intermédiaire de formules, telles que "pour tout x, MERLE(x) implique OISEAU(x)" ou "pour tout x, tout y et tout z, DONNER(x,y,z) implique POSSEDER(z,y)". Le premier exemple montre que l'on peut ainsi représenter des relations d'hyperonymie. Et de la même manière que le mécanisme d'héritage sur les liens "sorte-de" permet de déduire de nouvelles relations dans un réseau sémantique, c'est ici le mécanisme de base de la logique, l'**inférence**, qui joue : en effet, à partir des deux formules "pour tout x, MERLE(x) implique OISEAU(x)" et "pour tout x, OISEAU(x) implique VOLER(x)" on peut déduire la formule "pour tout x, MERLE(x) implique VOLER(x)".

On a vu qu'il était intéressant aussi de pouvoir **bloquer** l'héritage, pour exprimer par exemple qu'une *autruche* est un *oiseau* qui *ne vole pas*. La logique classique ne permet pas un tel traitement. Mais il existe des logiques, dites logiques **de défauts**, qui possèdent un tel mécanisme. En gros, cela revient à admettre, en plus des formules, des expressions un peu spéciales que l'on pourrait paraphraser par "on peut déduire qu'un oiseau vole si l'on n'a pas par ailleurs déjà déduit que cet oiseau ne vole pas".

Le deuxième exemple de formule que nous avons donné, et que l'on peut paraphraser par "le fait que x donne y à z implique que z possède y" appelle deux remarques. La première, c'est la **puissance d'expression** des formalismes logiques ; en effet, il n'est pas si facile d'exprimer dans un réseau sémantique une relation de ce genre (on peut le faire, bien entendu, et de différentes façons suivant les formalismes, mais aucune n'a cette simplicité). La deuxième, c'est une des limites de la logique classique : en effet, ce que l'on voudrait exprimer en fait, c'est que z possèdera y après que x le lui ait donné, c'est-à-dire une **dépendance temporelle**. Ceci n'est pas simple en logique classique : on peut toujours bien sûr ajouter à chaque prédicat un ou plusieurs arguments supplémentaires pour exprimer l'information temporelle, mais ce n'est qu'un pis-aller, dans la mesure où les lois temporelles ne sont pas intégrées au mécanisme de base de l'inférence. Il existe une autre solution : des extensions de la logique classique, les logiques **temporelles**, qui comportent des opérateurs interprétables par des expressions du genre "il existera un moment où il sera vrai que...", "il sera dorénavant toujours vrai que...", "il a toujours été vrai que...". Cela permet d'exprimer qu'une proposition Q ne sera vraie qu'après qu'une proposition P ait été vraie (à l'aide de combinaisons du type "il existera un moment où il sera vrai d'une part qu'il sera dorénavant toujours vrai que..., et d'autre part qu'il a toujours été faux que...").

On voit le parti que l'on peut tirer des formalismes logiques pour la représentation de connaissances lexicales. Mais c'est surtout à la représentation du sens d'une **phrase** que ces formalismes sont adaptés, grâce à l'interprétation naturelle d'une assertion par une formule prédicative. Une remarque terminologique s'impose ici, à propos de l'usage des termes "syntaxe" et

"sémantique" en linguistique et en logique, qui peut sembler à première vue un peu déroutante. Dans une théorie logique, une formule prédicative fait partie de la syntaxe : ce n'est que dans le cadre d'un modèle que l'on peut lui associer une **interprétation** sémantique. Ainsi, ce que l'on appelle le plus souvent en traitement automatique "représentation sémantique d'une phrase par la logique", c'est en fait la mise en correspondance de cette phrase avec une formule syntaxique d'un système logique...

Même si, comme on va le voir, la logique classique a des limites de ce point de vue, il faut d'abord souligner qu'elle permet d'exprimer aisément des relations de **quantification** qu'il est très difficile de représenter dans d'autres cadres. Ainsi prenons l'exemple de l'ambiguïté déjà citée (cf. *supra*, § 1.2.2.) : *Tous les élèves ont traité une question du problème.* Les deux interprétations possibles s'écrivent en logique (toujours en paraphrasant quelque peu...) :

1) "Il existe un x tel que l'on ait d'une part QUESTION(x) et d'autre part pour tout y, ELEVE(y) implique TRAITÉ (y,x)", qui traduit l'existence d'une question ayant la propriété d'avoir été faite par tous les élèves ;

2) "Pour tout y, ELEVE(y) implique qu'il existe un x tel que l'on ait QUESTION(x) et TRAITÉ (y,x)", qui traduit que pour chaque élève, il existe une question qui a été faite par cet élève.

Les logiques **typées** permettent de simplifier cette écriture (ce n'est bien sûr pas leur seul intérêt, loin s'en faut). En effet, dans ces logiques, on peut préciser le type d'une variable comme correspondant, dans une interprétation dans un monde donné, à un sous-ensemble des individus de ce monde. On peut alors préciser d'une part que x est de type "question" et y de type "élève" et les deux formules deviennent :

1) "Il existe un x tel que pour tout y on ait TRAITÉ (y,x)" ;

2) "Pour tout y, il existe un x tel que l'on ait TRAITÉ (y,x)".

La principale limite de la logique classique est bien connue : en assignant à chaque proposition une **valeur de vérité** (vrai ou faux), elles ne permettent pas de représenter des phrases dont la signification ne se laisse pas réduire à l'une de ces valeurs, soit parce qu'elles indiquent, explicitement ou non, une certaine forme d'indétermination, soit parce qu'elles conditionnent, par l'intermédiaire de modalités, l'attribution d'une valeur de vérité. Deux grandes familles de théories logiques peuvent pallier en partie ces difficultés.

Nous avons déjà parlé des logiques temporelles, qui permettent de prendre en compte le temps. De la même manière, on construit des logiques **modales** : deux opérateurs fondamentaux viennent compléter le dispositif de la logique classique. Suivant l'axiomatique précise que l'on choisit, on obtient des systèmes différents, correspondant plus ou moins bien à ce que l'on veut exprimer. Différentes paraphrases sont possibles pour ces opérateurs, chacune conduisant à l'expression de modalités précises. On peut ainsi les "traduire" par "il est

nécessaire que..." et "il est possible que...", ou par "il est obligatoire de ..." et "il est permis de..." ou encore par "un tel sait que..." et "un tel croit que...". Dans ces logiques (modales ou temporelles), la notion de modèle devient plus complexe : il faut en effet considérer un système de **mondes possibles**, reliés entre eux par une relation "d'accessibilité", pour interpréter les propositions "il est possible que..." par "il existe un monde où il est vrai que...", et "il est nécessaire que..." par "dans tous les mondes possibles, il est vrai que...". De plus, dans l'interprétation des logiques **épistémiques** ("un tel croit que..." et "un tel sait que..."), on est amené, pour rendre compte des croyances de plusieurs agents, à personnaliser pour chaque agent les relations entre les mondes possibles, créant ainsi ce que l'on appelle des univers de croyance (pour une théorie linguistique des univers de croyance, voir R. Martin 1987).

D'autres logiques cherchent à formaliser les notions d'indétermination en remplaçant l'ensemble des valeurs de vérité {vrai, faux} par des ensembles plus riches. Cet ensemble peut rester fini, avec trois valeurs {vrai, indéterminé, faux} ou plus : on obtient alors des logiques **multivaluées** qui se distinguent par le traitement des valeurs intermédiaires. Mais on peut aussi considérer un ensemble **continu** de valeurs de vérité : l'intervalle [0,1] par exemple ; on passe alors à ce que l'on appelle les logiques **floues**, qui, elles aussi, connaissent de nombreuses variantes.

On a donc un vaste choix de théories logiques dans lesquelles puiser pour construire un système d'analyse et de représentation sémantique. Ces théories ne sont pas en soi des théories de représentation sémantique : même si elles sont souvent motivées par des considérations linguistiques, elles ne se donnent pas comme objectif de traiter la sémantique des langues. En revanche, il existe des formalismes, à forte composante logique, qui ont été conçus **expressément** pour le traitement des langues. Deux d'entre eux se sont plus particulièrement imposés : la "théorie de la représentation du discours" ("Discourse Representation Theory", DRT en abrégé) de H. Kamp (1984) et la "grammaire universelle" de R. Montague (1970).

A chaque phrase, ou plus exactement à chaque proposition dans une phrase, Kamp associe une structure ("Discourse Representation Structure" ou DRS) qui est composée de deux parties : une en-tête où sont listées les variables qui représentent les différentes entités intervenant dans la phrase, et un corps qui contient des formules représentant les propriétés et relations qui "traduisent" le sens de la phrase. Ces structures peuvent être emboîtées, et les formules peuvent utiliser des variables d'une structure "emboîtante". Cela permet en particulier d'exprimer la co-référence par l'égalité de deux variables, à condition que les variables que l'on veut identifier soient accessibles dans la structure où l'on exprime la co-référence : cela donne donc des contraintes pour la résolution des anaphores, qui, sans être parfaites, permettent en tout cas de contourner un certain nombre d'obstacles que l'on rencontre en logique pour exprimer des relations simples de co-référence dans

des propositions hypothétiques, comme dans les célèbres "donkey sentences" (*Si Pedro a un âne, il le bat* ; *Tout fermier qui a un âne le bat*, etc.).

La théorie de R. Montague, quant à elle, comporte deux composantes : une grammaire catégorielle (cf. *supra*, chapitre 4 § 3.3.), traitant de la syntaxe, et un formalisme logique pour la sémantique. Ces deux composantes sont intimement liées : à chaque règle syntaxique correspond une règle sémantique et les constructions de la représentation syntaxique d'une phrase et de sa représentation sémantique obéissent à un parallélisme strict. La sémantique de R. Montague est fondée sur une **logique intensionnelle**. Ce qui caractérise une logique intensionnelle, c'est que termes et formules sont par définition fonction d'une famille de mondes (mondes possibles, mondes indexés par le temps, contextes...). Ainsi, si l'on s'en tient au temporel, un prédicat comme JEUNE( ) est associé à une fonction (valant vrai ou faux pour chaque individu) dépendant de l'indice *t* qui caractérise chaque monde : si l'on associe à JEAN un individu dans chaque monde (correspondant à la même personne évoluant au cours du temps), on peut rendre compte du fait qu'à la formule JEUNE(JEAN) est associée la valeur "vrai" dans certains de ces mondes et "faux" dans d'autres. Logique modale et logique temporelle sont donc naturellement intégrées à la logique intensionnelle. Sans aller plus loin dans cet aperçu (très simpliste) de la théorie de R. Montague, signalons qu'elle permet de traiter les problèmes d'opacité référentielle, évoqués plus haut au § 1.2.2.— pour un exposé de cette théorie, voir par exemple M. Chambreuil & J.C. Pariente (1990), F. Nef (1988) et surtout M. Galmiche (1991).

Signalons aussi les modèles de "grammaires applicatives" (cf. S. Shaumyan 1977 et J.P. Desclés 1990), qui s'appuient à la fois sur une grammaire catégorielle et sur une logique combinatoire typée. Enfin, il faut noter que réseaux sémantiques et formalismes logiques peuvent faire bon ménage : en fait, bien des systèmes aujourd'hui utilisent une approche **mixte**, dans laquelle on cherche à profiter des avantages des réseaux pour la représentation des connaissances et de la puissance d'expression qu'offre la logique. Pour ne citer qu'un exemple de ces systèmes que l'on appelle **hybrides**, le système VaDe développé au LIPN de l'Université Paris-XIII (cf. P. Grandemange 1992) utilise un réseau sémantique et une logique des défauts, pour implémenter une approche dite "à profondeur variable" (cf. *infra*, chapitre 8 § 2.3.).

3. Analyse sémantique : principes

Après avoir rappelé comment la plupart des systèmes mêlent étroitement syntaxe et sémantique, nous nous pencherons plus particulièrement sur la question du traitement de la polysémie, qui constitue l'une des pierres d'achoppement de l'analyse sémantique.

3.1. Relations entre analyse syntaxique et analyse sémantique

On peut concevoir une séparation complète de l'analyse syntaxique et de l'analyse sémantique (ce qui n'empêche pas, rappelons-le, que l'analyseur syntaxique utilise des connaissances sémantiques, en particulier des traits du type "±animé", pour mener à bien sa tâche). Dans ce cas, l'**entrée** de l'analyseur sémantique est l'**arbre syntaxique** associé à la phrase, et, en principe, un tel analyseur peut fonctionner avec des analyseurs syntaxiques de conception différente. C'est dans ce cadre que se placent des équipes qui ne s'intéressent qu'à l'analyse sémantique, et qui présupposent donc les problèmes syntaxiques résolus en amont de leur système. Mais en fait, la plupart des systèmes complets **mêlent** de façon étroite l'analyse syntaxique et l'analyse sémantique, tout au moins la partie de l'analyse sémantique qui consiste à calculer la structure prédicative.

C'est le cas d'abord des systèmes s'appuyant sur des théories qui supposent un **couplage** étroit entre syntaxe et sémantique : nous avons vu que la grammaire de R. Montague en faisait partie, en soutenant l'idée d'une correspondance "règle à règle" ("rule-to-rule hypothesis") entre une grammaire syntaxique catégorielle et une logique intensionnelle. Cela donne ce que l'on appelle une sémantique **compositionnelle** : en simplifiant, la "traduction" dans le langage de la logique s'opère noeud par noeud à partir de l'arbre syntaxique, des règles de composition permettant d'obtenir la valeur sémantique du noeud-père à partir de la valeur des noeuds-fils. Cette idée est largement répandue dans les théories qui utilisent l'unification (cf. *supra*, chapitre 4, § 4.4.). Ainsi la grammaire GPSG de G. Gazdar, dont nous n'avons présenté dans le chapitre 4 que l'aspect syntaxique, comporte aussi une composante sémantique, proche de la logique de R. Montague, et l'analyse sémantique est aussi couplée à l'analyse syntaxique : même s'il ne s'agit pas d'une correspondance règle à règle, mais plutôt structure à structure, elle obéit tout autant au principe de compositionnalité.

Dans d'autres systèmes, l'analyse sémantique peut même **guider** l'analyse syntaxique. Ainsi dans le système FRUMP conçu par G. DeJong (1982) pour analyser de petits textes en utilisant la théorie de R. Schank, c'est la présence de certains mots (typiquement les verbes) qui permet de sélectionner des primitives conceptuelles qui guident ensuite le traitement : l'analyseur cherche alors à rem-

plir directement les rôles sémantiques prévus par une telle primitive en s'appuyant sur les connaissances syntaxiques qui sont attachées à la primitive sélectionnée. Dans ce type de système, il n'y a donc à aucun moment de construction d'une représentation syntaxique de la phrase, même pas dans une phase transitoire. On peut aussi citer l'analyseur par "experts-mots" ("word expert parsing") de S. Small & C. Rieger (1982), dans lequel les connaissances syntaxiques et sémantiques sont attachées à chaque morphème (lexical ou grammatical) : le contrôle de l'analyse passe de mot en mot, par un mécanisme très proche de celui des "tableaux noirs" (cf *supra*, chapitre 3 § 4.4.).

3.2. Le traitement de la polysémie

Comme on l'a vu tout au long de ce chapitre, la polysémie (aussi bien lexicale que grammaticale) constitue un des problèmes majeurs auquels sont confrontés les analyseurs sémantiques. En effet, ce qui caractérise une unité polysémique, c'est que sa signification dépend de la phrase dans laquelle elle est insérée. La compositionnalité conduit donc à un cercle vicieux : pour calculer le sens de la phrase, il faut partir du sens de chacun de ses éléments, et pour avoir le sens de chaque élément, il faut connaître la phrase entière. Dans les approches classiques, deux solutions sont possibles pour briser ce cercle vicieux. La première consiste à traiter toute polysémie comme une ambiguïté locale : on essaie donc autant d'hypothèses qu'il y a d'acceptions du mot, et l'on ne garde que la solution qui satisfait à toutes les contraintes de compositionnalité dans la phrase donnée (la ou les solutions: il peut y en avoir plusieurs en cas d'ambiguïté réelle, globale). Cette solution n'est en fait pas très viable : il faut se contenter de descriptions très grossières des unités polysémiques si l'on ne veut pas se heurter à l'explosion combinatoire. L'autre solution consiste à ajouter aux règles de compositionnalité des règles de "recatégorisation", qui permettent de modifier la signification d'un noeud en fonction d'un noeud de niveau supérieur : on confère alors à une unité polysémique une signification "première" (la plus fréquente, par exemple), mais au cours du calcul, cette signification peut être amenée à changer sous l'effet d'une interaction avec ses voisins. Cette solution est aussi très lourde : les règles de recatégorisation présentent une combinatoire très élévée, puisque des éléments fort éloignés dans la phrase peuvent jouer un rôle décisif dans l'interprétation d'un polysème.

Un certain nombre de recherches consacrées spécifiquement à ce problème avancent des solutions dans lesquelles l'interaction entre le sens local d'une unité et le sens global de la phrase est mieux prise en compte. On peut citer, entre autres, les travaux de G. Hirst (1988) qui propose un système d'analyse dans lequel chaque unité est progressivement instanciée au fur à mesure du traitement de la phrase : il appelle cette technique "les mots polaroïds", le sens de chaque mot se révélant progressivement à la manière d'une photographie à développement instantané. Une autre approche consiste à utiliser des réseaux connexionnistes pour rendre compte de la relation entre les divers éléments co-

textuels (dans la phrase) et les différentes acceptions d'un mot : cf. C. Harris (1990) et C. Fuchs & B. Victorri (eds.) (1988) ; voir aussi J. Veronis & N. Ide (1990) pour un autre type d'utilisation de réseaux connexionnistes.

4. Perspectives

Comme on a pu le voir tout au long de ce chapitre, les approches de la sémantique effectuées en vue de traitements automatiques ont eu, jusqu'ici, pour objectif essentiel la recherche de **formalismes** de représentation (formalismes issus de la logique ainsi que des réseaux sémantiques). Ce faisant, on a considéré comme allant de soi que la sémantique se concevait comme une **interprétation** de la syntaxe (abordée dans une perspective compositionnelle) combinée à une sémantique lexicale très rudimentaire. Et l'on a assez largement fait comme si la description linguistique du fonctionnement sémantique de la langue ne posait pas de problème : on a en particulier laissé de côté des pans entiers de la sémantique linguistique (comme par exemple le secteur de l'énonciation et de la référence).

Pourtant, il semble que se fasse jour à l'heure actuelle tout un courant de recherches qui tend à dénoncer l'**inadéquation** des modèles formels proposés, et à redécouvrir une certaine tradition européenne de description linguistique (en particulier sur le fonctionnement sémantique des **catégories grammaticales** comme le temps, l'aspect, la modalité, la détermination, etc.).

Très révélatrice à cet égard est la prise de conscience, de la part d'informaticiens travaillant dans le domaine, que la non-réductibilité de la sémantique linguistique à une sémantique formelle, loin de constituer un "défaut" des langues, en fait au contraire la richesse et l'intérêt : d'où une attention toute particulière portée à des phénomènes sémantiques spécifiques aux langues "naturelles", comme l'ambiguïté, la polysémie, la paraphrase, etc.

Par ailleurs, si les réseaux sémantiques et la logique ont fourni jusqu'ici l'essentiel des formalismes utilisés dans les systèmes d'analyse sémantique, d'autres approches explorent actuellement des voies radicalement différentes. En particulier, on assiste aujourd'hui à un regain d'intérêt pour des représentations de type **topologique** et **géométrique**, qui font le lien entre perception et langage, sous l'influence, entre autres, des grammaires cognitives américaines. Ces approches n'ont pas encore donné naissance à des formalismes opérationnels, mais elles constituent dès à présent une voie extrêmement prometteuse pour dépasser le type de difficultés auxquelles se heurtent les représentations logiques.

Catherine FUCHS et Bernard VICTORRI
(ELSAP-CNRS)

Repères bibliographiques

1. Présentations d'ensemble :

CARRE, R. *& al.* (1991) : *Langage humain et machine*, Paris, Presses du CNRS.

[Le ch. 1 : "L'étude et le traitement automatique des langues naturelles" et le ch. 3 : "Le traitement de l'écrit" contiennent, entre autres, quelques développements assez généraux sur les questions d'analyse sémantique automatique.]

COULON, D. & KAYSER, D. (1986) : Informatique et langage naturel : présentation générale des méthodes d'interprétation des textes écrits, *Technique et Science Informatiques*, 5 : 2, Paris, Gauthier-Villars, 103-128.

[Article de synthèse ; sur l'analyse sémantique, voir pp. 110-115.]

KAYSER, D. (1985) : Des machines qui comprennent notre langue, *La Recherche*, 16 : 170, Paris, 1198-1212.

[Article de vulgarisation consacré aux problèmes de la compréhension automatique ; sur l'analyse sémantique, voir pp. 1203 *sq.*]

PITRAT, J. (1985) : *Textes, ordinateurs et compréhension*, Paris, Eyrolles.

[Un ouvrage de base très accessible et très vivant.]

SABAH, G. (1988) : *L'intelligence artificielle et le langage* ; vol. I : *Représentations des connaissances ;* Paris, Hermès.

[Voir notamment le ch. 3 : "Grammaires de cas" ; le ch. 6 : "Logique formelle" ; le ch. 7 : "Réseaux sémantiques" ; et le ch. 8 : "Primitives et représentations procédurales".]

SMITH G.W. (1991) : *Computers and Human Language*, Oxford, Oxford University Press.

[Voir notamment les ch. 4 ("Structure and search") et 8 ("Lexical semantics") pour le lexique et les réseaux sémantiques, ainsi que les ch. 9 ("Phrase and sentence semantics") et 10 ("Integrating syntactic and semantic processing") sur la sémantique de la phrase.]

2. Modèles linguistiques :

FUCHS, C. & LE GOFFIC, P. (1992) : *Les linguistiques contemporaines ; repères théoriques*, Paris, Hachette.

[Pour une introduction aux grands courants actuels en sémantique linguistique, voir le ch. 10 : "Sémantique, logique et cognition", le ch. 11 : "Enonciation et pragmatique", et le ch. 12 : "La théorie des opérations énonciatives de A. Culioli".]

- Sémantique componentielle :

LE NY, J-F. (1979) : *La Sémantique psychologique*, Paris, P.U.F.

[Pour un prolongement psycho-linguistique de la sémantique componentielle, voir le ch. IV : "L'organisation componentielle et la nature des sèmes".]

RASTIER, F. (1987) : *Sémantique interprétative*, Paris, P.U.F.

[Eaboration d'une sémantique componentielle, qui débouche sur l'interprétation des textes.]

- Grammaires de cas :

FILLMORE, Ch. (1968) : The case for case, dans E. Bach & R. Harms (eds.) : *Universals in linguistic theory*, Chicago, Holt Rinehart & Winston, 1-90.

[L'article historique de référence, dont se sont inspirés maints traitements automatiques.]

POTTIER, B. (1992) : *Théorie et analyse en linguistique*, Paris, Hachette.

[Pour une autre approche linguistique des cas, voir le ch. X : "Le système casuel".]

- Sémantique générative :

GALMICHE, M. (1975) : *Sémantique générative*, Paris, Larousse.

[Introduction synthétique à ce courant de sémantique linguistique qui, à la fin des année 60, fut une dissidence de la grammaire chomskienne orthodoxe (en s'opposant à la sémantique "interprétative") et qui proposait des décompositions sémantiques des unités lexicales.]

- Grammaires cognitives :

LAKOFF, G. (1987) : *Women, fire and dangerous things : what categories reveal about the mind*, Chicago, University of Chicago Press.

[Ouvrage de l'un des chefs de file américains du courant de la linguistique cognitive.]

LANGACKER, R. (1987 / 1991) : *Foundations of cognitive grammar*, vol. I : *Theoretical prerequisites* / vol. II : *Descriptive application* Stanford, Stanford University Press.

[Théorie de la signification comme processus cognitivo-linguistique de structuration et de symbolisation de contenus conceptuels ; et applications.]

TALMY, L. (1988) : Force dynamics in language and cognition, *Cognitive Science*, 12 : 1, Norwood N.J., Ablex, 49-100.

[Une approche cognitive du langage centrée sur la notion de dynamique.]

- Sémantique énonciative :

CULIOLI, A. (1990): *Pour une linguistique de l'énonciation*, vol. I : *Opérations et représentations*, Ophrys.

[Présentation d'une théorie des opérations linguistiques constitutives de la construction de l'énoncé, et illustration sur diverses catégories de langue.]

- Sémantiques de la référence :

KLEIBER, G. (1990) : *L'article "le" générique : la généricité sur le mode massif*, Genève, Droz.

[Un exemple d'analyse sémantique d'un marqueur linguistique de la référence, non réductible à un opérateur logique classique.]

MARTIN, R. (1987) : *Langage et croyance: les "univers de croyance" dans la théorie sémantique*, Bruxelles, Mardaga.

[Exposé d'une théorie de sémantique linguistique qui recourt à la notion d'"univers de croyance" ; illustration sur de nombreux exemples du français.]

3. Formalismes de représentation :

- Décompositions et réseaux sémantiques :

BRACHMAN, R. & SCHMOLZE, J. (1985) : An overview of the KL-ONE knowledge representation system, *Cognitive Science*, 9 : 2, Norwood N.J., Ablex, 171-216.

[Présentation du système KL-ONE.]

DESCLES, J-P. (1987) : "Réseaux sémantiques", *Langages*, 87, Paris, Larousse, 55-78.

[Réflexion critique sur le statut théorique des réseaux sémantiques.]

FAHLMAN, S. (1979) : *NETL: a system for representing and using real-world knowledge*, Cambridge Mass., M.I.T. Press.

[Présentation d'un système recourant à des réseaux "à propagation de marqueurs".]

HENDRIX, G. (1979) : Encoding knowledge in partitioned networks, dans N. Findler (ed.) : *Associative networks : representation and use of knowledge by computers*, New-York, Academic Press, 51-92.

[Présentation des réseaux "partitionnés".]

QUILLIAN, R. (1968) : Semantic memory, dans M. Minsky (ed.) : *Semantic information processing*, Cambridge Mass., M.I.T. Press, 227-270.

[Article historique introduisant les réseaux sémantiques pour rendre compte des phénomènes de mémorisation.]

RASTIER, F. (1991) : *Sémantique et recherches cognitives*, Paris, P.U.F.

[Pour une présentation critique du recours aux réseaux en traitement automatique de la sémantique par l'intelligence artificielle, voir le ch. 4: "La sémantique des réseaux" et le ch. 5 : "Formalismes de l'intelligence artificielle et représentations du signifié lexical".]

SCHANK, R. (1972) : Conceptual dependency : a theory of natural language understanding, *Cognitive Psychology*, III : 4, New-York, Academic Press, 552-631.

[Présentation du formalisme des "dépendances conceptuelles".]

SCHANK, R. & ABELSON, R. (1977) : *Scripts, plans, goals and understanding*, Hillsdale, Lawrence Erlbaum.

[L'ouvrage de référence présentant les "scénarios".]

WILKS, Y. (1975) : Preference semantics, dans E. Keenan (ed.) : *Formal semantics of natural language*, Cambridge, Cambridge University Press, 329-348.

SOWA, J. (1984) : *Conceptual structures: information processing in man and machine*, Reading Mass., Addison Wesley.

[Présentation de la théorie des "graphes conceptuels".]

SOWA, J. (ed.) (1991) : *Principles of semantic networks. Explorations in the representation of knowledge*, San Mateo, Morgan Kaufmann.

[Série d'articles des principaux spécialistes des réseaux sémantiques, qui fait le point sur la recherche dans ce domaine.]

WILKS, Y. (1978) : Making preferences more active, *Artificial Intelligence*, 11, Amsterdam, Elsevier, 197-223.

[Deux articles consacrés à la "sémantique préférentielle", qui assouplit le recours aux traits sémantiques.]

- Formalismes logiques :

Introductions générales aux logiques :

AUDUREAU, E. *& al.* (1989) : *Logique temporelle ; sémantique et validation de programmes parallèles*, Paris, Masson.

[Voir le ch 1 : "Eléments de logique modale et temporelle" et le ch. 2 : "Notions élémentaires de logique temporelle"]

CRESWELL, M. (1988) : *Semantical essays ; possible worlds and their rivals*, Dordrecht, Kluwer.

[Introduction aux "mondes possibles", aux "logiques des situations et des attitudes" et aux problèmes de quantification et de référence].

KAYSER, D. (1990) : Adéquation et inadéquation de la logique au traitement sémantique des langues, *Modèles Linguistiques*, XII : 1, Lille, 119-136.

[Réflexion sur ce que peut apporter la logique à l'analyse sémantique: inadéquation de la logique classique, et plaidoyer en faveur d'une logique non-monotone.]

LEA SOMBE (1988) : "Inférence non classique en intelligence artificielle", *Actes des journées nationales du P.R.C. "I.A."*, Teknea.

[Etude comparative des différents points de vue sur l'inférence non monotone.]

THAYSE, A. *& al.* (1988 / 1989 / 1990) : *Approche logique de l'intelligence artificielle*, 4 volumes, Paris, Dunod.

[Ouvrage très bien documenté donnant les bases de la logique classique et de certaines logiques utilisées en intelligence artificielle (logiques modales, logiques non monotones, etc.)]

Grammaire universelle de R. Montague et sémantique intensionnelle :

CHAMBREUIL, M. & PARIENTE, J.C. (1990) : *Langue naturelle et logique ; la sémantique intensionnelle de Richard Montague*, Berne, Lang.

[Présentation assez technique de la grammaire universelle de R. Montague.]

GALMICHE, M. (1991) : *Sémantique linguistique et logique ; un exemple : la théorie de R. Montague*, Paris, P.U.F.

[Présentation plus linguistique, illustrée sur des exemples du français.]

MONTAGUE, R. (1970) : English as a formal language, repris dans R. Thomason : *Formal philosophy, selected papers of Richard Montague*, New-Haven, Yale University Press.

[Article de référence de l'auteur.]

Logiques intensionnelles, logiques modales, logiques de la référence :

BORILLO, A. *& al.* (1982) : *Approches formelles de la sémantique naturelle*, Travaux en Informatique Logique Linguistique, Toulouse, L.S.I.

[Recueil d'articles sur la formalisation de phénomènes sémantiques linguistiques.]

KULAS, J. *& al.* (eds.) (1988) : *Philosophy, language and artificial intelligence*, Dordrecht, Kluwer.

[Voir en particulier la Partie II : "Semantic aspects of natural language", la Partie II : "Connecting syntax with semantics", la Partie IV : "Natural language and logical form" et la Partie V : "Possible-worlds and situation semantics".]

NEF, F. (ed.) (1983) : "La sémantique logique ; problèmes d'histoire et de méthode", *Histoire, Epistémologie, Langage*, 5 : 2, Lille, Presses Universitaires.

[Recueil de contributions sur la logique intensionnelle et ses prolongements.]

NEF, F. (1988) : *Logique et langage : essais de sémantique intensionnelle*, Paris, Hermès.

[Présentation de la sémantique de R. Montague, ainsi que des développements ultérieurs de la sémantique intensionnelle et des théories logiques de la référence.]

KAMP, H. (1984) : A theory of truth and semantic representation, dans J. Groenendijk *& al.* (eds.) : *Truth, interpretation and information*, Dordrecht, Foris.

[Texte de base présentant la DRT.]

Grammaire applicative :

DESCLES, J.P. (1990) : *Langages applicatifs, langues naturelles et cognition*, Paris, Hermès

[Présentation de la "grammaire applicative et cognitive".]

SHAUMYAN, S. (1977) : *Applicational grammar as semantic theory of natural language*, Chicago University Press.

[Présentation du modèle de la "grammaire applicative universelle".]

4. Traitement informatiques :

DEJONG, G. (1982) : An overview of the FRUMP system, dans G. Lehnert & M. Ringle (eds.) : *Strategies for natural language processing*, Hillsdale, Erlbaum, 149-176.

[Un analyseur fondé sur les primitives conceptuelles de R. Schank.]

[Voir également R. Schank & R. Abelson (1977) : cf. *supra*, § 3.]

FUCHS, C. ET VICTORRI, B. (eds) (1988) : "Vers un traitement automatique de la polysémie grammaticale", *T.A. Informations*, 29, Paris, Klincksieck.

[Un traitement de la polysémie de marqueurs grammaticaux français à l'aide de réseaux connexionnistes.]

HARRIS, C (1990) : Connectionism and cognitive linguistics, *Connection Science*, 2 : 1 / 2, Corfax, 7-33. Repris dans N. Sharkey (ed.) (1992) : *Connectionnist natural language processing*, Dordrecht, Kluwer Academic Press, 1-27.

[Utilisation d'un réseau connexionniste pour modéliser la polysémie de *over*.]

HIRST, G. (1988) : Resolving lexical ambiguity computationally with spreading activation and polaroid words, dans S. Small *& al.* (eds.) (1988), 73-107.

[Présentation des "mots polaroïds".]

GRANDEMANGE, P. (1992) : Raisonnement à profondeur variable: le système VaDe, Rapport interne du L.I.P.N.-Paris Nord.

[Exemple de système hybride utilisant à la fois un réseau sémantique et un formalisme logique.]

SMALL, S. & RIEGER, C. (1982) : Parsing and comprehending with word experts (a theory and its realization), dans G. Lehnert & M. Ringle (eds.): *Strategies for natural language processing*, Hillsdale, Erlbaum, 89-147.

[Un système d'analyse à contrôle distribué ("experts-mots").]

SMALL, S.*& al.* (eds) (1988) : *Lexical ambiguity resolution*, San Mateo, Morgan Kaufmann.

[Dans la partie 1 ("Computer models"), série d'articles consacrés à différentes méthodes de traitement de la polysémie et de l'homonymie.]

VERONIS, J. & IDE, N. (1990) : Word sense disambiguation with very large neural networks extracted from machine readable dictionaries, *Actes du colloque COLING 90*, Helsinki, vol. 2, 389-394.

[Une tentative d'utilisation de réseaux connexionnistes de très grande taille, confectionnés automatiquement à partir de dictionnaires d'usage, pour lever les ambiguïtés lexicales]

DEUXIÈME PARTIE

LES DOMAINES DES TRAITEMENTS AUTOMATIQUES DES LANGUES

6 TRAITEMENT DE LA PAROLE

Après un historique rapide du traitement automatique de la parole (tant en synthèse qu'en reconnaissance), nous présenterons ses divers domaines en insistant plus particulièrement sur les stratégies de synthèse et de reconnaissance qui mettent en jeu la manipulation de modèles et d'outils linguistiques. Au cours de ce chapitre, nous faisons référence à des mécanismes physiques et articulatoires supposés connus, et rappelés dans le chapitre 1 de la première partie (cf. *supra*).

1. Repères historiques

Comme dans tous les autres domaines du traitement automatique, les différentes étapes historiques qui vont être évoquées à propos du traitement de la parole correspondent non seulement à l'évolution des préoccupations scientifiques, mais aussi à des nécessités pratiques, à des enjeux économiques et parfois même à des décisions politiques et stratégiques.

1.1. Regard historique sur la synthèse de la parole

Nous distinguerons trois phases : la première (ou préhistoire), que nous évoquerons à titre anecdotique, s'amorce dès la Renaissance et s'étend jusqu'au début du XX^{e} siècle. La seconde période (dès les années 30) correspond à des préoccupations scientifiques (meilleure connaissance des mécanismes de production et de perception de la parole) ; ces préoccupations vont de pair avec des impératifs économiques : dans le domaine des télécommunications, il faut exploiter la redondance du signal de parole en compressant ce signal (réduction

de débit), ceci afin d'augmenter la capacité de transport des lignes téléphoniques déjà installées. Cette période est le point-clé de l'histoire de la synthèse puisqu'elle correspond au développement de techniques de **codage** de la parole (vocodeurs à canaux) et à la mise au point de **synthétiseurs** : les synthétiseurs à formants et les synthétiseurs articulatoires. Enfin, dès les années 1970, grâce à la diffusion de l'informatique, de nouveaux axes de recherche sont exploités, de nouvelles **techniques de codage** (la prédiction linéaire notamment) voient le jour. Par ailleurs, outre leur rôle d'outil expérimental, certains systèmes de synthèse développés ont également une finalité commerciale.

On situe en général les débuts de la parole artificielle au XVII^e siècle ; pourtant dès la fin de la Renaissance les premiers modèles physiques de production du signal de parole sont développés : F. Bacon déclare dans *New Atlantis* : "Nous produisons encore à volonté des sons articulés et toutes les lettres de l'alphabet, soit les consonnes, soit les voyelles que nous imitons, ainsi que les différentes espèces de voix et de chants des animaux terrestres et des oiseaux". Vers 1630, M. Mersenne propose un projet de dispositif parlant proche de l'orgue à tuyaux : les voyelles proviennent de tuyaux à embouchure de flûte et divers dispositifs sont décrits pour imiter les consonnes.

Au Siècle des Lumières, cet engouement pour la parole artificielle persiste : dans les années 1780 apparaît en Russie un exemple de simulation du conduit vocal dont la réalisation est attribuée à un certain Kratzeinstein ; composée d'un ensemble de résonateurs acoustiques excités par une anche vibrante, la machine produit cinq voyelles. En 1791, une autre machine parlante est inventée par le baron Von Kempelen, gentilhomme de la cour d'Autriche-Hongrie, considéré aujourd'hui par les experts en parole comme "le véritable pionnier de la synthèse de la parole" (cf. J.S. Liénard 1977). Sa machine comprend un soufflet et une chambre à air comprimée munie d'une anche vibrante. Un résonateur constitué d'un cuir déformable à la main est utilisé pour produire des sons voisés. En outre, les consonnes sont crées par fermeture de certains orifices ou par des sifflets actionnés par des leviers. Cette machine peut émettre une vingtaine de sons différents.

Au XIX^e siècle, quelques autres machines sont construites sur les principes de la machine de Von Kempelen, dont celle de J. Faber présentée à Vienne en 1835 (cf. J.S. Liénard 1977).

Mais ce n'est véritablement qu'en 1939, quand apparaît au laboratoire Bell le premier codeur de voix électrique, le VODER ("Voice Operation Demonstrator"), que la synthèse est née (cf. Calliope 1989) ; sa fonction première est d'étudier le rôle relatif des différentes composantes du signal détectées par l'analyse acoustique.

En 1950, à l'issue d'études nombreuses sur la production et la transmission de la parole, le "Pattern Play-Back", relecteur de spectrogrammes, est développé aux Etats-Unis (laboratoire d'Haskins). Il joue un rôle essentiel pour la compréhension des caractéristiques acoustiques et articulatoires de la parole.

En 1953, une nouvelle technique de synthèse, fondée sur la simulation articulatoire du conduit vocal, est présentée aux Etats-Unis (au MIT). Un analogue du conduit vocal est réalisé non plus par des filtres modulant le spectre de la source mais par une simulation directe de la géométrie du conduit vocal (fonction d'aire).

Enfin, l'anné 1959 est marquée par la présentation d'une nouvelle technique de synthèse : deux **synthétiseurs à formants** sont réalisés aux Universités de Stockholm et d'Edimbourg. L'enveloppe spectrale d'un signal de parole est décrite par ses composantes essentielles : les formants, qui correspondent aux fréquences de résonance des cavités du conduit vocal (cf. *supra,* chapitre 2).

Les années 1970 marquent la phase la plus récente de l'histoire de la synthèse, elle est liée au développement des théories sur le traitement du signal numérique et à l'explosion des ordinateurs. Deux voies de recherche, déjà connues, sont explorées : reproduire le signal de parole à partir de simulations fonctionnelles du conduit vocal humain (la synthèse se fait ici par formants), ou bien simuler la propagation de l'onde sonore dans le conduit vocal à partir de connaissances physiologiques, articulatoires et mécaniques (il s'agit alors de modélisation articulatoire).

1.2. Regard historique sur la reconnaissance de la parole

On considère en général quatre grandes périodes dans l'histoire de la reconnaissance. La première (dans les années 50), où on utilise les outils de l'électronique analogique, correspond à la réalisation de systèmes de reconnaissance d'unités élémentaires (mots, syllabes, phonèmes, traits) prononcées le plus souvent par un seul sujet. La seconde période (dès les années 60) est marquée par l'utilisation de l'ordinateur qui s'impose au fur et à mesure des années, la taille du vocabulaire traité et le nombre de locuteurs testés augmentent progressivement, les premiers systèmes de reconnaissance de la parole continue voient le jour. Dans les années 1970, on tente d'utiliser explicitement les niveaux dits supérieurs (syntaxique et sémantique) pour améliorer les performances de reconnaissance malgré les erreurs de décodage acoustico-phonétique. A la fin de la décennie 70-80, la recherche est de nouveau axée sur le problème crucial de décodage acoustico-phonétique (cf. Calliope 1989).

En 1952, les recherches du Suisse Dreyfus Graf aboutissent à la réalisation d'un système permettant de segmenter l'onde sonore en phonèmes. Ce système, le **phonétographe**, est perfectionné au cours des années, et en 1961 une application de dictée est présentée : la machine écrit les lettres de l'alphabet soigneusement prononcées par l'inventeur. Toujours en 1952, l'invention du premier système de reconnaissance des chiffres aux Etats-Unis est le fait de chercheurs du laboratoire Bell. Ce système reconnaît les chiffres prononcés isolement par un locuteur donné, après apprentissage préalable. En 1956,

la première machine à écrire phonétique est présentée aux Etats-Unis, ce système reconnaît 10 syllabes prononcées isolément par un seul locuteur. En 1959, les recherches effectuées au laboratoire Lincoln aux Etats-Unis aboutissent aux premières réalisations de systèmes informatiques : un programme permet la reconnaissance de dix voyelles différentes de lexèmes monosyllabiques.

En 1962, les chercheurs d'IBM aux Etats-Unis mettent au point un prototype permettant de reconnaître des chiffres isolés à partir d'une segmentation en consonnes et en voyelles. En 1965, à l'université de Kyoto, le premier système de reconnaissance de parole continue est réalisé pour le japonais à partir d'une segmentation en unités phonétiques, des contraintes sur les séquences de suites de trois phonèmes ("triphones") sont utilisées.

En 1970, à Grenoble, le Français J-P. Tubach utilise explicitement les contraintes syntaxiques et sémantiques pour assister la reconnaissance des niveaux inférieurs, dans le cadre du développement d'un système de dictée vocale en mots isolés (cf. J-P. Tubach 1970). En 1972, le premier système de reconnaissance de parole est commercialisé aux Etats-Unis : la machine à reconnaître de Treshold (le VP100) est un système capable de reconnaître 30 mots après apprentissage avec un taux d'erreur de 1 à 2%.

Mais c'est surtout le projet extrêmement ambitieux ARPA/SUR ("Advanced Research Projects Agency / Speech Understanding Research"), lancé par le département de la Défense aux Etats-Unis, qui marque la décennie 70-80. Il s'agit de développer des systèmes capables de comprendre des phrases prononcées de façon continue avec un vocabulaire de 1000 mots en utilisant des contraintes linguistiques et en appliquant des méthodes de type intelligence artificielle. (Une analyse exhaustive de ce projet est présentée dans D. Klatt 1977).

Dès la fin des années 1970, du fait des limites des systèmes développés, le décodage acoustico-phonétique, considéré comme le problème majeur de la reconnaissance, est à nouveau l'axe prioritaire des recherches en reconnaissance. Trois approches sont envisagées : la première est fondée sur le traitement du signal et la reconnaissance des formes ; la seconde se fonde sur une approche de type intelligence artificielle (développement de systèmes experts en lecture de spectrogrammes) ; la dernière, toujours très fructueuse aujourd'hui, est axée sur une approche probabiliste du décodage acoustico-phonétique (modélisation markovienne).

2. Les systèmes de synthèse de la parole

L'objectif d'un système de synthèse de la parole est de produire un énoncé oral à partir d'une représentation phonétique de celui-ci. Deux méthodes peuvent être utilisées : la synthèse par concaténation d'éléments pré-enregistrés (il s'agit de **codage**) et la synthèse de **vocabulaire illimité**, qui peut se faire soit à partir d'une entrée graphémique (c'est le cas de la **synthèse à partir du texte**), soit à partir de concepts sémantiques (il s'agit alors de **synthèse par concepts**). Le choix d'une de ces méthodes dépend de l'application visée, de la qualité de la synthèse exigée et du coût de revient du système de synthèse développé.

2.1. La synthèse par mots

Nous ne nous attarderons pas sur cette méthode de synthèse, qui, certes, est la plus employée aujourd'hui dans les applications grand public, mais qui présente peu d'intérêt pour la recherche fondamentale et ne concerne pas la linguistique.

Un système de synthèse par mots remplit la même fonction qu'un magnétophone digital (enregistrement d'une séquence de parole prononcée par un locuteur humain, codage et compression de celle-ci, enfin restitution du message). Dans la restitution du signal de parole, le vocabulaire est limité à celui qui a été prononcé et la voix est invariable (celle du locuteur enregistré). Cette technique présente l'avantage d'être économique et facile à mettre en oeuvre, ce qui explique sa diffusion dans des applications diverses (jouets électroniques, appareils électro-ménagers, serveurs vocaux, etc.). Néanmoins, elle a l'inconvénient de ne permettre la modification d'un message qu'en faisant appel à la personne qui a prononcé ce message. Si celle-ci n'est plus disponible, un ré-enregistrement complet est nécessaire ; ceci n'est évidemment pas envisageable pour des applications de grande envergure. Par ailleurs, la technique de codage employée (faible, moyen ou élevé) conditionne la qualité de la synthèse qui n'est pas toujours optimale. Enfin, la prosodie restreinte aux unités enregistrées limite les performances de tels systèmes.

2.2. La synthèse de vocabulaire illimité

Nous définirons la synthèse de vocabulaire illimité comme la production par un ordinateur d'un énoncé oral de longueur quelconque qui n'a jamais été prononcé auparavant. Pour ce faire, plusieurs étapes sont à distinguer (cf. Figure I).

La première étape, décrite dans les chapitres 1 et 2 (cf. *supra*), est de nature linguistique. Dans un premier temps, il s'agit de faire correspondre à

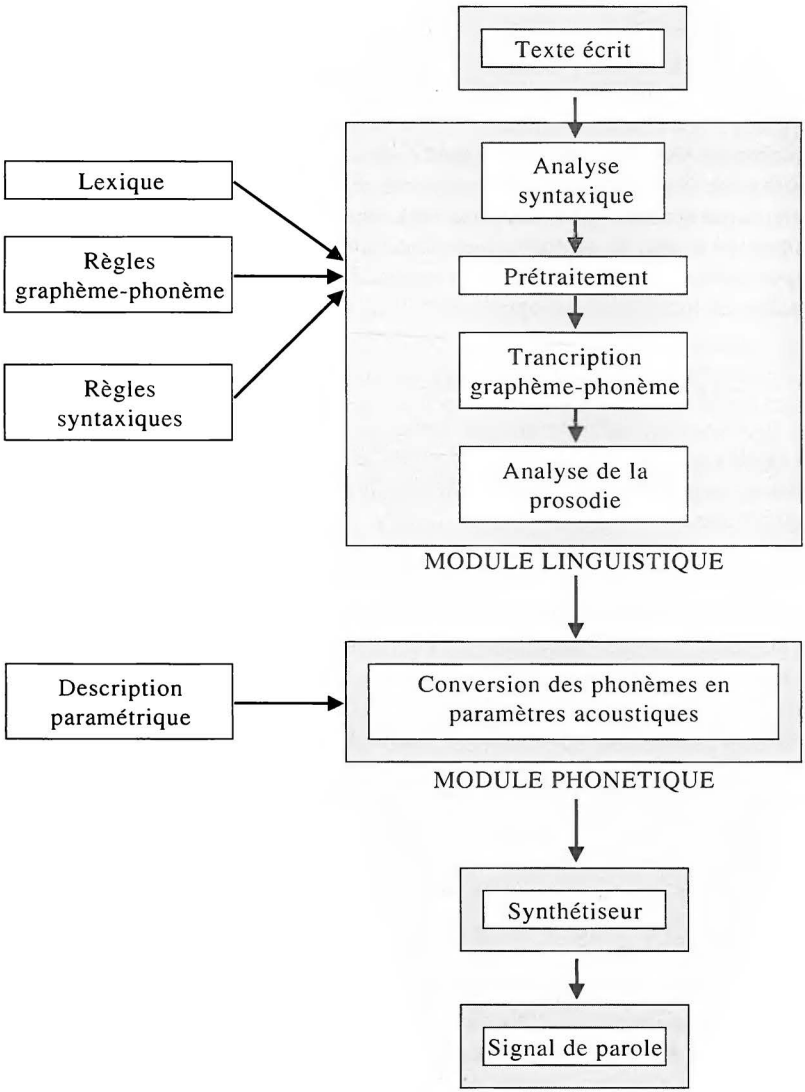


FIGURE I
Architecture générale d'un système de synthèse à partir du texte (synthétiseur à formants)

un texte écrit d'une langue donnée une suite de symboles phonémiques. On parle alors de **transcription graphème-phonème**. Le développement d'un **modèle prosodique** pour la génération automatique de la prosodie (génération de la durée, allocation de pauses, d'accents, attribution de contours prosodiques qui traduisent les variations de la courbe mélodique) est également nécessaire. Pour ce faire, une analyse syntaxique du texte permet de détecter les différentes parties du discours, leur fonction dans l'énoncé, enfin d'identifier la structure syntaxique de la phrase à synthétiser et sa modalité.

La deuxième étape, décrite dans le chapitre 1 (cf. *supra*), est de nature phonétique. Elle consiste à transformer une suite de symboles phonétiques discrets en un signal de parole continue, intelligible et, autant que faire se peut, naturelle. Il s'agit donc d'effectuer la transition entre une représentation linguistique relativement abstraite et la réalité acoustique.

La dernière étape, présentée ci-dessous, est relative au développement du synthétiseur utilisé pour produire la parole synthétique.

2.3. Les techniques de synthèse

Différents synthétiseurs peuvent être utilisés pour produire de la parole synthétique. Nous présentons ici les quatre techniques de synthèse les plus classiques : les vocodeurs à canaux (qui ont connu leur temps de gloire, mais qui, on verra pourquoi, ne sont plus utilisés aujourd'hui), le codage par prédiction linéaire, enfin, les synthétiseurs à formants et les synthétiseurs articulatoires.

2.3.1. Les vocodeurs à canaux

La fonction d'un vocodeur à canaux, tel qu'il a été développé dans les années 1940 (codeur de Dudley), était, rappelons-le, de servir d'outil d'analyse pour étudier les mécanismes fondamentaux de production et de perception de parole. En retour, une connaissance accrue des processus de la phonation devait permettre d'améliorer le fonctionnement des synthétiseurs.

Un tel système est donc composé de **deux modules** : un module d'analyse et un module de synthèse.

Le module d'**analyse** assume deux fonctions : un détecteur de mélodie détermine les variations de la fréquence fondamentale (fréquence de vibration des cordes vocales), une analyse spectrale du signal de parole renseigne sur le caractère voisé ou non des sons. Cette analyse est réalisée par un banc de filtres passe-bande (canaux fréquentiels dont la fréquence est comprise dans un certain intervalle). L'analyseur fournit donc à chaque instant une image quantifiée du spectre de parole et de son évolution.

Le module de **synthèse** doit produire un signal sonore dont le spectre en fréquence se rapproche au mieux du modèle fourni par l'analyseur. Le message parlé est reconstitué selon le processus inverse de l'analyse. Le synthétiseur est muni d'une source de signaux périodiques dont on modifie le spectre

artificiellement. Un canal supplémentaire fournit un signal indiquant la fréquence fondamentale associée aux sons voisés. Le signal ainsi obtenu possède un spectre qui se rapproche du spectre de la parole originelle.

Cette technique de synthèse a fait l'objet d'une quantité de réalisations ; néanmoins l'**intelligibilité** de la parole synthétique est toujours restée insuffisante. Les spécialistes du domaine attribuent ce semi-échec à quatre raisons majeures (cf. J-S. Liénard 1977) : d'abord, il est nécessaire de quantifier le squelette phonétique plus finement quant à la fréquence des formants ; deuxièmement la distinction entre sons voisés et sons non voisés n'est pas toujours tranchée dans la parole réelle ; par ailleurs, la détection de la mélodie par simple filtrage de la fréquence fondamentale est insuffisante ; enfin, il semble qu'il soit impossible de réaliser une synthèse de qualité sans prendre en compte les caractéristiques du conduit vocal.

De fait, cette technique a été délaissée au profit de méthodes de codage plus adéquates, le codage par "prédiction linéaire" notamment. Les méthodes de synthèse qui utilisent des modèles "fonctionnels" (synthétiseur à formant ou synthétiseur articulatoire) peuvent également être utilisées.

2.3.2. Codage par prédiction linéaire

Rappelons que les systèmes de synthèse dits "par concaténation d'unités" font appel à deux processus différents : un processus de décodage et un processus de concaténation. Dans la méthode de prédiction linéaire (en anglais LPC : "linear predictive coding"), ces deux processus sont effectués simultanément, les unités acoustiques étant directement décodées et synthétisées avec leurs valeurs prosodiques cibles. Cette méthode se fonde sur les connaissances de la production de la parole supposée linéaire de la source au conduit vocal (cf. G. Fant 1960). La source, associée à un train d'impulsions périodiques pour les sons voisés ou à un bruit blanc pour les sons non voisés, est excitée par un filtre dont la fonction de transfert représente le conduit vocal. L'ensemble du conduit vocal est vu comme un système dynamique linéaire, modélisable suivant trois axes : un modèle glottal (source), un modèle du conduit vocal qui évalue sa fonction de transfert (amplification des harmoniques dans le conduit vocal), et un modèle du conduit nasal. A mi-chemin entre la méthode spectrale et la méthode temporelle, la prédiction linéaire permet de mesurer la fonction de transfert du filtre. Cette technique repose sur l'exploitation de la redondance existant dans la forme du signal de parole : l'échantillon actuel Sn du signal à l'instant *n* peut être calculé à partir de la somme pondérée des échantillons passés (Sn-1, Sn-2..., Sn-p), 'p' correspondant au nombre de valeurs à prendre en considération. Un algorithme calcule les coefficients de cette combinaison, 8 à 16 pour obtenir une synthèse de qualité acceptable (plus il y en a, plus l'analyse est fine car moins la différence entre le signal originel et le signal prédit est grande). L'utilisation d'une fenêtre temporelle glissante de taille fixe permet de connaître l'évolution du signal de parole à chaque instant.

Cette technique présente de nombreux avantages (analyse automatique, simplicité de mise en oeuvre, fidélité au timbre originel, etc.). Cependant une représentation trop simpliste de la source d'excitation semble être à l'origine d'une qualité bruyante de la parole synthétique. En outre, le modèle est mal adapté à la représentation de certains sons tels que les nasales ou les fricatives voisées. Différentes améliorations ont été apportées à la méthode pour limiter l'aspect bruyant de la voix et la sensation métallique associée à la parole synthétique. Pour une revue des algorithmes de codage appliqués à la synthèse à partir du texte, on pourra consulter E. Moulines (1990).

2.3.3. Synthétiseurs à formants

Par son principe même (recherche d'analogie entre l'appareil électrique et le conduit vocal), le synthétiseur à formants simule mieux que le vocodeur le fonctionnement de l'appareil vocal. L'organe de commande est aujourd'hui un ordinateur qui analyse l'évolution des formants dans un certain nombre de séquences prononcées par un locuteur humain et détermine les paramètres correspondants qui sont ensuite stockés dans une bibliothèque ; chaque formant est codé par sa fréquence centrale, son amplitude et sa largeur de bande. L'information perceptive concernant la nature des sons voisés est fournie par les formants : à un ensemble de filtres résonants est associée une courbe en fréquence qui reproduit celle du conduit vocal. Chacun de ces filtres amplifie une bande de fréquence correspondant à un formant déterminé. Les signaux issus d'une impulsion périodique ou d'une source de bruit sont également modélisés. La synthèse des sons voisés nécessite l'utilisation de plusieurs circuits de formants ; les uns, commandés, déterminent la fréquence des trois premiers formants, les autres, fixes, sont relatifs à la nasalité. La synthèse des sons non voisés fait intervenir une source de bruit associée à deux filtres résonants ; des paramètres permettent de définir l'amplitude de la source et la fréquence des deux formants de bruit.

Si le principe de tels synthétiseurs est séduisant (les paramètres utilisés sont étroitement corrélés à la production et à la propagation du signal de parole dans le conduit vocal), il ne tient pas compte de la mobilité des articulateurs dans le conduit vocal, source de grande variabilité (cf. *supra,* chapitre 1).

2.3.4. Synthétiseurs articulatoires

Tandis que les méthodes de synthèse à formants reposent sur l'utilisation de modèles perceptifs fonctionnels (il s'agit de produire un signal de parole dont le spectre en fréquence est proche de la parole humaine entendue), l'objectif est ici différent : il faut simuler le fonctionnement physique du conduit vocal en tenant compte de la mobilité des organes phonatoires (langue, mâchoires, lèvres, etc.). On s'intéresse donc aux mécanismes de production de la parole et à la variabilité articulatoire. Les paramètres de commande mis en jeu dans de tels synthétiseurs sont associés à la pression subglottique, à la tension des cordes vocales et à la position relative des arti-

culateurs dans le conduit vocal. On peut distinguer deux types de simulateurs du conduit vocal : pour les premiers, dont le fonctionnement est **statique**, il est impossible de modifier automatiquement les paramètres représentant la forme du conduit vocal ; le fonctionnement **dynamique** des seconds permet de faire varier automatiquement la forme de l'appareil vocal artificiel.

Cette modélisation, qui tend à se rapprocher le plus possible de la réalité articulatoire, est séduisante. Cependant, les données articulatoires dont disposent les chercheurs sont encore insuffisantes pour pouvoir définir finement les paramètres de commande du simulateur et pour aboutir à des modélisations en tout point satisfaisantes. La fonction d'aire notamment, qui permet de préciser certaines caractéristiques articulatoires comme la position de la langue ou l'ouverture de la bouche, est difficile à déterminer. Enfin, la source est très difficile à modéliser. De fait, de tels synthétiseurs restent du domaine de la recherche fondamentale, même si de très bons résultats préliminaires ont pu être constatés.

3. Les systèmes de reconnaissance automatique de la parole

La fonction d'un système de reconnaissance automatique de la parole est de fournir une réponse à un message prononcé par un humain. Cette réponse peut être donnée par écrit (**dictée vocale**), elle peut être vocale (**dialogue homme-machine**), ou elle peut correspondre à une action (**commande vocale**). Pour qu'elle satisfasse aux exigences de l'utilisateur, cette réponse suppose une bonne "compréhension" du message prononcé. Or, la variabilité rencontrée en parole (cf. *supra,* chapitre 2) rend les processus de décodage et de compréhension complexes.

De même qu'il est nécessaire de différencier les méthodes de synthèse en fonction des techniques mises en œuvre, des objectifs visés et des résultats obtenus, différents systèmes de reconnaissance doivent être distingués. Les critères généralement utilisés pour effectuer cette distinction sont le degré de complexité de la tâche de reconnaissance et le mode de reconnaissance choisi (global ou analytique).

3.1. Degré de complexité de la tâche de reconnaissance

Evaluer la complexité d'une tâche de reconnaissance suppose de poser les questions suivantes : la reconnaissance est-elle monolocuteur, multilocuteur ou indépendante du locuteur ? s'agit-il de reconnaissance de mots isolés ou de parole continue ? la reconnaissance se fait-elle sur un grand vocabulaire ou sur un vocabulaire limité ? quelles sont les contraintes imposées à l'utilisateur ?

3.1.1. Reconnaissance monolocuteur, multilocuteur ou indépendante du locuteur

Trois modes de reconnaissance, présentés par ordre de complexité croissante, peuvent être utilisés :

- En reconnaissance **monolocuteur** la complexité de la tâche est limitée au fait que la voix d'un seul sujet peut être reconnue après apprentissage préalable (l'utilisateur du système de reconnaissance doit prononcer une ou plusieurs fois les unités utilisées lors de la reconnaissance, des mots par exemple, ces unités, stockées en mémoire, constituent l'image acoustique de référence à laquelle sera comparée l'unité à reconnaître). Mais déjà ici la difficulté est réelle, puisque la reconnaissance doit passer outre les problèmes engendrés par la **variabilité intra-locuteur** (prononciation variable d'une séquence de parole par un locuteur donné).

- La reconnaissance **multilocuteur** nécessite également une phase d'apprentissage préalable. Il s'agit ici non plus de reconnaître un seul locuteur mais plusieurs ; la difficulté est donc accrue, puisqu'il faut résoudre les problèmes inhérents à la **variabilité inter-locuteur** (prononciation variable d'une séquence de parole par des locuteurs distincts).

- La reconnaissance **indépendante du locuteur** permet de traiter le vocabulaire à reconnaître de façon standard, sans faire d'ajustement pour les nouveaux utilisateurs du système de reconnaissance.

3.1.2. Reconnaissance de mots isolés ou de parole continue

En reconnaissance de **mots isolés**, les mots prononcés par le ou les locuteurs à reconnaître sont séparés par une pause. En revanche, dans un processus de reconnaissance de **parole continue**, il est nécessaire de tenir compte des phénomènes de co-articulation en frontière de mots, phénomènes qui mettent en jeu des connaissances phonétiques et linguistiques complexes.

3.1.3. Le vocabulaire à reconnaître

S'il est vrai que plus la taille du vocabulaire augmente, plus la reconnaissance est difficile, la nature du vocabulaire à reconnaître est également à prendre en compte : il est plus difficile de reconnaître des mots phonétiquement proches que différents, des mots courts que des mots longs.

3.1.4. Contraintes imposées à l'utilisateur

Moins les contraintes imposées à l'utilisateur sont importantes, plus la reconnaissance s'avère complexe. Outre les points évoqués précédemment (apprentissage préalable, prononciation en mots isolés, nature et taille du vocabulaire à reconnaître), les contraintes liées à la syntaxe utilisée plus ou moins rigide, au mode d'élocution (articulation soignée, débit lent) et à l'environnement acoustique (tolérance aux bruits parasites plus ou moins forte) constituent autant de moyens pour contourner, ou du moins limiter les

difficultés de la reconnaissance. Aujourd'hui, on ne sait toujours pas traiter la parole naturelle en milieu bruité.

3.2. Mode de reconnaissance : global ou analytique

On distingue deux classes de systèmes de reconnaissance : les systèmes dits **auto-organisateurs** utilisent des méthodes mathématiques pour effectuer la reconnaissance (programmation dynamique et modélisation markovienne par exemple) ; les systèmes **fondés sur des connaissances** utilisent explicitement les connaissances des experts humains en matière d'acoustique, de phonétique et de linguistique.

3.2.1. Reconnaissance globale

Il y a peu de temps encore, on pouvait distinguer deux approches en reconnaissance : la première fondée sur la comparaison directe de l'image acoustique des mots à reconnaître avec une image de référence, il s'agissait de **reconnaissance globale** ; la seconde, fondée sur l'identification d'unités phonétiques, on la désignait par le terme **reconnaissance analytique**. En reconnaissance globale, les objectifs étaient de développer des systèmes pouvant reconnaître un petit vocabulaire, prononcé mot à mot par un seul locuteur. Au minimum une phase d'apprentissage préalable à la reconnaissance imposait au locuteur de prononcer les différents mots du vocabulaire à reconnaître. La (ou les) image(s) acoustique(s) de ces mots étai(en)t stockée(s) en mémoire ; lors de la reconnaissance l'image acoustique des mots à reconnaître était comparée aux images acoustiques des mots de référence, avec choix du plus proche voisin et rejet en cas de ressemblance trop incertaine. Si l'approche reste globalement la même aujourd'hui, des modifications importantes ont été introduites quant à l'unité de reconnaissance - ou **unité de décision** - choisie, qui n'est pas nécessairement le mot mais se rapproche de plus en plus du phonème, afin de prendre en compte les phénomènes de co-articulation entre les sons, et de manière à rendre l'unité de reconnaissance moins tributaire du vocabulaire à reconnaître. La reconnaissance de phrases complètes, composées de mots du vocabulaire connus par la machine et prononcées en continu, est également une approche de reconnaissance globale. Des progrès sont aussi à noter quant au nombre de locuteurs pouvant être reconnus : des systèmes multilocuteurs de reconnaissance globale de mots isolés ont été développés, voire même commercialisés.

Ainsi, les frontières entre reconnaissance globale et reconnaissance analytique ne sont plus aussi nettes qu'elles ont pu l'être : l'unité de décision et le mode de reconnaissance choisis (mots isolés / phrases, monolocuteur / multilocuteur) ne peuvent plus servir de critères pour distinguer ces deux modes de reconnaissance. Des différences notables persistent malgré tout : d'une part, en reconnaissance globale, la reconnaissance de la parole s'effectue toujours à

partir du niveau acoustique, d'autre part, les connaissances mises en œuvre dans le décodage sont implicites et limitées, les outils mathématiques (comparaison dynamique, quantification vectorielle par exemple) étant les principaux garants d'une bonne reconnaissance. En revanche, la reconnaissance analytique utilise des connaissances linguistiques, c'est elle qui est mise en œuvre dans des systèmes de compréhension de la parole continue. Différents outils peuvent être utilisés pour intégrer ces connaissances (outils fournis par l'intelligence artificielle et connaissances d'experts par exemple). Cette méthode de reconnaissance nous intéresse directement puisqu'on est amené à y manipuler de façon plus ou moins prégnante des formalismes linguistiques.

3.2.2. Systèmes de compréhension de la parole continue

Rappelons que le terme de "compréhension de la parole" ("speech understanding") a été introduit par les chercheurs américains du projet ARPA pour désigner des systèmes qui utilisent des informations linguistiques afin de compenser les limites du décodage acoustico-phonétique. La distinction reconnaissance / compréhension est essentiellement opératoire : "si le terme de compréhension se justifie, c'est essentiellement parce que la finalité des systèmes proposés est de provoquer une action correspondante à la demande orale de départ en utilisant l'ensemble des informations liées au langage et à l'univers de l'application traitée" (cf. Calliope 1989).

Les critères que nous utiliserons pour distinguer les systèmes de compréhension parlée sont les suivants : le niveau de langage traité et la stratégie de reconnaissance.

Quatre grands **types de traitements** mis en œuvre dans les systèmes de compréhension parlée peuvent être distingués :

- La reconnaissance par **mots clés** ("word spotting") consiste à donner l'illusion à l'utilisateur qu'il est en présence d'un système "intelligent" reconnaissant la parole naturelle, alors qu'il ne s'agit que d'un système "stupide" de détection de mots. Nous ne détaillerons pas ce type de reconnaissance, qui peut convenir pour des applications restreintes, mais où la communication est réduite à sa plus simple expression, et qui ne fait pas intervenir de véritables processus de compréhension.

- Les systèmes **guidés par la syntaxe** utilisent la syntaxe pour guider la compréhension. L'analyse syntaxique permet d'effectuer le passage d'un treillis de phonèmes à une phrase reconnue. Elle peut être effectuée de deux façons : en limitant à chaque pas de la reconnaissance le nombre de mots à tester ou en sélectionnant dans un treillis de mots reconnus la ou les phrases syntaxique(s) correcte(s). Cette stratégie de reconnaissance est essentiellement utilisée dans des systèmes de compréhension qui reposent sur l'utilisation de langages artificiels à vocabulaire restreint.

- La reconnaissance de **langage quasi-naturel** consiste à utiliser la langue naturelle avec un minimum de restrictions, afin de limiter la phase d'appren-

tissage du langage. Parmi les premiers systèmes expérimentaux utilisant un tel type de reconnaissance, on peut citer en France, le système Myrtille II (cf. J.M. Pierrel 1981) réalisé au CRIN à Nancy. De tels systèmes nécessitent la mise en œuvre d'une procédure de dialogue complexe et doivent prendre en compte plusieurs niveaux de connaissances linguistiques (pragmatique, sémantique, syntaxique, lexicale et, autant que faire se peut, prosodique).

- La **dictée vocale** ne tolère aucune restriction quant au langage employé. Seules les contraintes relatives au mode d'élocution peuvent être utilisées pour diminuer la difficulté. Si les études menées dans ce domaine sont nom-

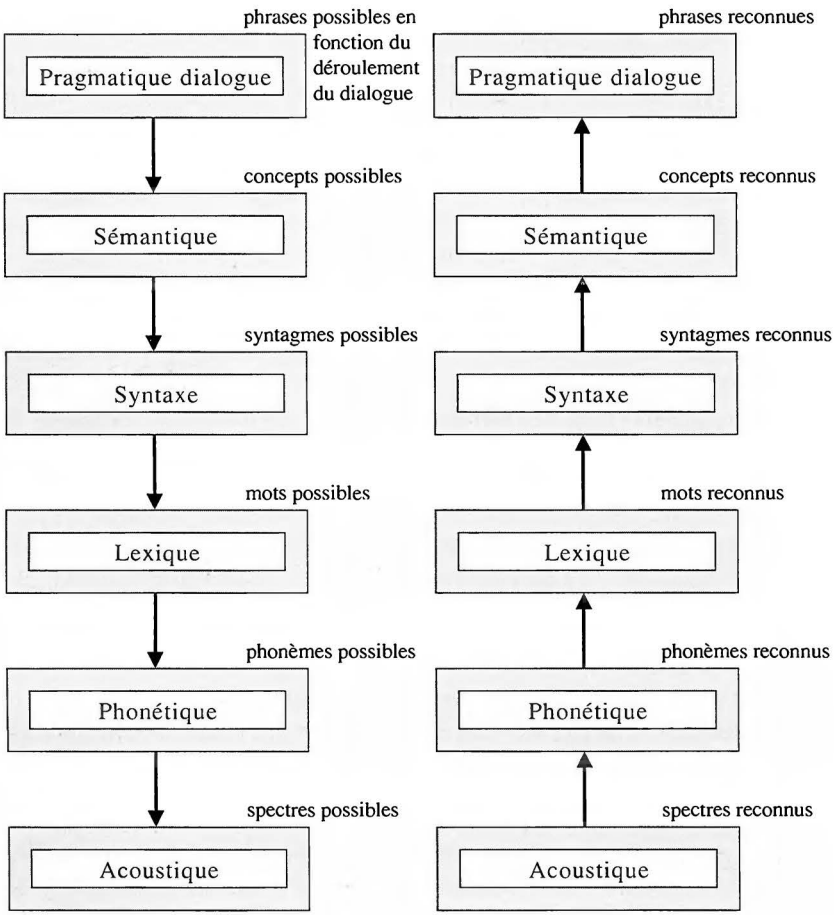


FIGURE II
Stratégies de reconnaissance possibles en reconnaissance analytique

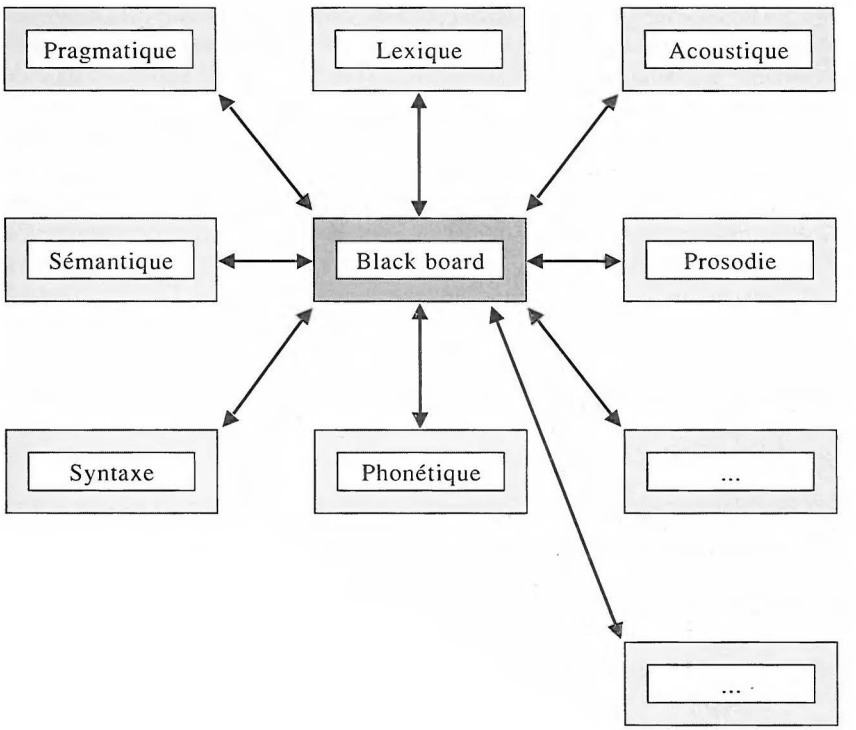


FIGURE III
Représentation simplifiée du modèle "tableau noir" ("blackboard")

breuses aujourd'hui, rappelons que la dictée vocale automatique reste un idéal extrêmement difficile à atteindre.

Pour accéder aux connaissances acoustiques et linguistiques, quatre **stratégies** peuvent être envisagées (cf. Figures II et III). Quelle que soit la stratégie employée, les niveaux d'analyse sont globalement les mêmes, seul l'**ordre** de traitement diffère :

- Dans une stratégie dite **ascendante**, il s'agit de reconstituer une phrase à partir des niveaux dits inférieurs (acoustique et phonétique) en prenant en compte les informations lexicales, syntaxiques, sémantiques et pragmatiques jusqu'à la compréhension du message.

- Dans une stratégie **descendante**, les contraintes imposées par les niveaux supérieurs permettent de générer des hypothèses lexicales qui doivent permettre de remédier aux déficiences des niveaux inférieurs et notamment aux erreurs de décodage acoustico-phonétique. Si ces hypothèses ne sont pas confirmées au niveau du signal reçu, elles sont remises en cause ; dans le cas contraire, elles sont acceptées et on en informe les niveaux supérieurs.

- Une stratégie mixte **ascendante-descendante** permet de confronter les thèses venant des niveaux supérieurs aux hypothèses proposées par l'analyse acoustico-phonétique. Le système français Myrtille II utilise un tel type de stratégie.

Dans les systèmes développés aux Etats-Unis dans les années 70 (projet ARPA-SUR), des stratégies plus complexes sont employées. Celles-ci ne reposent plus comme précédemment sur l'organisation hiérarchique des différents niveaux, mais sur une structure dite de **tableau noir** ("blackboard"). Dans de tels systèmes, un niveau peut échanger des informations avec tout autre niveau par l'intermédiaire d'un tableau noir central, où l'information est disponible jusqu'à ce qu'un autre niveau utilise cette information.

4. Perspectives

Divers problèmes fondamentaux restent à résoudre pour accroître les performances des systèmes de traitement automatique de la parole. Selon les experts, un effort soutenu doit se poursuivre dans trois directions complémentaires au moins :

- Des théories unifiées devraient permettre d'améliorer les modèles de **génération** automatique de la **prosodie** en synthèse de la parole et d'utiliser à bon escient les paramètres prosodiques comme indices robustes en reconnaissance, ce qui est encore loin d'être le cas aujourd'hui.

- En reconnaissance des efforts doivent se poursuivre en **décodage acoustico-phonétique** qui, comme nous l'avons montré au chapitre 1 (cf. *supra*), reste la principale pierre d'achoppement des chercheurs.

- Enfin, des connaissances inter-disciplinaires (linguistiques, psychologiques, acoustiques, etc.) accrues devraient permettre d'être plus à même de concevoir des **architectures** adaptées au traitement automatique de la parole.

Nous aimerions conclure ce chapitre en présentant les perspectives de recherche offertes par le développement de systèmes de communication **multimodale**, qui, depuis la fin des années 80, s'inscrit dans le paysage du traitement automatique de la parole.

L'objectif d'un système de communication multimodale est de permettre à un sujet humain de communiquer avec un système informatique en utilisant les modalités de communication qui lui conviennent (parole, geste, clavier, souris, gant tactile, etc.). Le système informatique, quant à lui, doit pouvoir restituer une information à l'aide de modalités complémentaires, voire redondantes. Pour ce faire, deux axes de recherche doivent être explorés : l'identification des concepts multimodaux et le développement de techniques multimodales. En linguistique, ces objectifs ouvrent des perspectives de

recherche, extrêmement intéressantes sur deux points au moins : la mise en place d'outils d'analyse multimodaux et de modèles linguistiques tenant compte du geste, de la parole et de la langue écrite, et le recueil de données d'observations, l'identification et l'analyse de corpus en vue de définir un modèle de l'utilisateur à la fois statique et dynamique.

Dans le domaine informatique, le développement de gestionnaires multimodaux, l'étude d'architectures logicielles multimodales permettant l'utilisation simultanée de plusieurs modalités dans un univers dynamique, enfin, la mise en place de postes d'observation multimodaux, sont également nécessaires.

Cette nouvelle discipline suscite un intérêt réel des chercheurs du domaine, la preuve en est le lancement en France, dès 1990, d'une opération de recherche concertée sur le thème "interface homme-machine multimodale" (cf. J. Coutaz et J. Caelen 1991).

Anne LACHERET-DUJOUR
(Université de Caen, LIMSI-CNRS et ELSAP-CNRS)

Repères bibliographiques

TRAITEMENT AUTOMATIQUE DE LA PAROLE

1. Ouvrages généraux :

CALLIOPE (1989) : *La parole et son traitement automatique,* Paris, Masson. [(cf. *supra,* chapitre 1).]

CARRÉ, R. *& al.* (1991) : *Langage humain et machine*, Paris, Presses du CNRS. [voir le ch. 4 : "Le traitement de la parole".]

LIENARD, J.S. (1977) : *Les processus de la communication parlée,* Paris, Masson.

[Synthèse accessible aux non spécialistes des recherches en communication parlée depuis leurs origines, appuyée sur des expérimentations concrètes et de nombreuses références bibliographiques.]

2. Synthèse de la parole :

FANT, G. (1960) : *Acoustic Theory of Speech Production,* The Netherland, Mouton's-Gravenhage.

[Ouvrage de référence en phonétique articulatoire tant sur le plan théorique que dans une perspective de validation pratique (application à la synthèse) sur les mécanismes de production de la parole.]

KLATT, D. (1987) : Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America,* 82 : 3, New York, Acoustical Society of America, 737-793.

[Historique de la synthèse à partir du texte depuis ses débuts.]

MOULINES, E. (1990) : *Algorithmes de codage et de modification des paramètres prosodiques pour la synthèse de la parole à partir du texte,* Thèse de Doctorat, ENST, Paris.

[Thèse présentant de façon critique les différents algorithmes de codage de la parole appliqués à la synthèse à partir du texte.]

3. Reconnaissance de la parole :

KLATT, D. (1977) : Review of the ARPA Speech Understanding Project, *Journal of the Acoustical Society of America,* 62, New York, Acoustical Society of America, 1345-1366.

[Revue détaillée des recherches menées sur la reconnaissance de la parole aux Etats-Unis durant la décennie 70-80.]

LEA, W.A. (1980) : *Trends in Speech Recognition,* New Jersey, Englewood Cliffs, Prentice-Hall.

[(cf. *supra,* chapitre 1).]

PIERREL, J-M. (1981) : *Etude et mise en oeuvre de contraintes linguistiques en compréhension automatique du discours continu,* Thèse d'Etat, Université de Nancy I.

[Présentation de travaux portant sur le développement d'un système de dialogue oral homme-machine en français considéré comme une référence par les spécialistes.]

TUBACH, J-P. (1970) : *Reconnaissance automatique de la parole ; étude et réalisation fondée sur les niveaux acoustique, morphologique, syntaxique,* Thèse d'Etat, Université de Grenoble.

[Thèse présentant une des premières réalisations d'un système de reconnaissance de la parole continue en français fondé sur des contraintes non seulement acoustiques mais également linguistiques.]

COMMUNICATION MULTIMODALE

COUTAZ, J. & CAELEN, J. (1991) : L'opération de recherche concertée interface homme-machine multimodale, *Deuxièmes Journées nationales du GRECO-PRC Communication Homme-Machine*, Toulouse.

[Synthèse claire sur les recherches en communication multimodale dans une perspective de dialogue oral homme-machine. Les concepts fondateurs de la communication multimodale y sont présentés et discutés.]

TAYLOR, M. N. *& al.* (1986) : *Structure of multilodal dialog*, Amsterdam, North Holland, Coll. "Human factors in information".

[Etat de l'art et perspectives en matière de dialogue multimodal.]

7
Traduction Automatique

Avec le traitement de la parole, la traduction représente sans doute l'un des plus anciens domaines des traitements automatiques des langues : dès la fin de la Seconde Guerre mondiale, des enjeux directement politiques conduisaient en effet les Etats-Unis à s'intéresser à ce secteur. Cette relative ancienneté explique l'existence de systèmes développés et commercialisés en nombre plus important que dans des domaines plus récents, comme la compréhension automatique ou la génération automatique de textes (cf. *infra*, chapitres 8 et 9). Ajoutons que la pression des milieux économiques et politiques, sensible dès le début des recherches dans le domaine, demeure très présente et tend souvent à reléguer au second plan les problématiques purement théoriques.

Après avoir brièvement retracé l'historique des recherches dans le domaine (§ 1.), et présenté schématiquement les types de produits existants et les types de besoins en traduction qu'ils visent à satisfaire (§ 2.), nous étudierons les différentes architectures possibles des systèmes de traduction automatique (§ 3.), ainsi que les traitements effectués par les différents modules constitutifs de ces systèmes (§ 4.), en insistant sur les problèmes linguistiques soulevés par ces traitements.

1. Repères historiques

On peut schématiquement distinguer trois grandes étapes dans l'histoire de la traduction automatique : la première (de la fin des années 40 jusqu'en 1965, date du rapport ALPAC), a été marquée successivement par l'enthou-

siasme des débuts, puis par un certain désenchantement ; la seconde (de 1965 à 1975) a été une période de relative stagnation, due aux contrecoups du rapport ALPAC ; la troisième enfin (de 1975 à nos jours) témoigne d'une vitalité nouvelle, notamment en matière de recherche et développement.

1.1. Des débuts (fin des années 40) au rapport ALPAC (1965) : l'enthousiasme et le désenchantement

Vers la fin des années 40 aux Etats-Unis, W. Weaver propose d'utiliser les techniques du déchiffrage cryptographique pour traduire des textes de façon automatique, et en 1952 se tient au M.I.T. la première conférence sur la traduction automatique : c'est le tout début des recherches dans le domaine. L'enthousiasme avec lequel furent accueillis les premiers essais s'explique dans le contexte économico-politique de cette période de l'immédiat après-guerre : c'est l'époque de la rivalité entre les Etats-Unis et l'URSS (guerre froide et course à l'espace) ; aussi les premiers systèmes pilotes aux Etats-Unis (comme le G.A.T. — "Georgetown Automatic Translation" — élaboré à l'Université de Georgetown par l'équipe de B. Dorstert et R. Mc Donald) furent-ils consacrés à la traduction de textes scientifiques du russe vers l'anglais. Les espoirs suscités par ces premières recherches furent assez vite déçus : on avait très largement sous-estimé les difficultés théoriques de l'entreprise. Aussi dès le tournant des années 60, une relative désillusion succéda-t-elle à l'enthousiasme des débuts : certains commencèrent à s'interroger sur la faisabilité de traductions entièrement automatiques de bonne qualité, dénonçant de surcroît le climat de compétition et la précipitation qui présidaient aux recherches dans le domaine (cf. Y. Bar-Hillel 1960).

Cette désillusion s'explique en partie par la nature des systèmes de cette période, qui participaient de ce que l'on a par la suite appelé des systèmes de "première génération". Ils effectuaient en effet des traductions **mot à mot** à partir d'un dictionnaire sans construire de représentation syntaxique de la phrase et ils adoptaient une approche dite **directe** (cf. *infra*, § 3.1.), ne procédant pas à une analyse de la langue-source séparée de la génération de la langue-cible. Disons-le brutalement : aucune considération de nature linguistique n'avait présidé à l'élaboration de ces systèmes, et les spécialistes du domaine (mathématiciens et informaticiens) n'ont compris la nécessité de réfléchir sur la langue que lorsqu'ils ont été confrontés concrètement à des échecs.

Contrairement à cet empirisme naïf, qui caractérisait les premiers travaux américains, les recherches menées à la même époque en URSS présentaient une tendance nettement plus théorique : intégrant les questions d'analyse linguistique, elles s'orientaient déjà vers l'élaboration d'un "langage intermédiaire" universel (ne participant ni de la langue-source ni de la langue-cible) en termes de modélisations logico-mathématiques : cf. les travaux de l'équipe

de V. Rosenzweig à Moscou, à laquelle ont participé des linguistes de renom (comme par exemple I. Mel'chuk). Cette tendance préfigurait les systèmes ultérieurs dits de "deuxième génération" (voir *infra*, § 1.2.).

C'est en 1965, alors que commençait à retomber l'enthousiasme des débuts, que la commission **ALPAC** ("Automatic Language Processing Advisory Committee") présentait le rapport qui lui avait été demandé par les agences gouvernementales finançant depuis quelque dix ans aux Etats-Unis les travaux de la vingtaine d'équipes de recherche en traduction automatique. Ayant examiné les résultats des traductions russe-anglais des systèmes américains de première génération, le rapport critiquait la qualité et le coût de ces traductions (qui nécessitaient d'importantes révisions par l'humain), et s'interrogeait sur l'utilité de poursuivre un tel programme de recherche.

Les conséquences de ce rapport ne se sont pas fait attendre : les crédits des groupes de recherche sur la traduction automatique ont été considérablement réduits, voire supprimés, tant aux Etats-Unis que dans le reste du monde. Aux yeux des décideurs, le rapport concluait purement et simplement que la traduction automatique était vouée à l'échec ; en réalité, le rapport dénonçait surtout les insuffisances d'un certain type de systèmes, et il insistait sur la nécessité de développer les recherches en linguistique informatique — condition nécessaire pour progresser dans le domaine.

1.2. De 1965 à 1975 : la stagnation

Le brusque déclin des sources de financement a considérablement réduit le nombre et l'importance des équipes de recherche pendant les dix années qui ont suivi le rapport ALPAC. Certaines ont néanmoins réussi à poursuivre leurs travaux : citons aux Etats-Unis les travaux sur les systèmes METAL (à l'Université du Texas) et SYSTRAN (Latsec Inc.), au Canada les travaux sur TAUM (à Montréal), et en France les recherches au CETA — devenu en 1976 le GETA, sous la direction de B. Vauquois (à Grenoble), principal centre public français de recherche dans le domaine.

La plupart de ces systèmes commençaient à s'orienter vers ce que l'on a appelé des systèmes de "deuxième génération". Les principales caractéristiques que l'on s'accorde généralement à reconnaître à un système de deuxième génération sont les suivantes :

- une approche dite **indirecte** (cf. *infra*, § 3.1.), où la traduction opère sur la base de deux modules distincts d'analyse de la langue-source et de génération de la langue-cible ;

- une stricte **séparation** entre les connaissances linguistiques (grammaires et dictionnaires) et la partie logicielle (programmes) ;

- une traduction qui se fait à un niveau plus **profond** que dans les systèmes de première génération : l'analyse passe par une phase syntaxique (qui

pourra ultérieurement être enrichie par des éléments d'analyse sémantique, voire même contextuelle) ;

- des objectifs plus restreints en matière de **domaine** de textes à traduire : ces systèmes visent surtout la traduction de textes techniques appartenant à un domaine bien délimité (comme par exemple TAUM-METEO sur les bulletins météorologiques) — le vocabulaire et les constructions sont ainsi plus facilement répertoriés, et les risques d'ambiguïtés moindres ;

- enfin une **diversification** des langues traduites : on sort très largement du couple anglais-russe des débuts, et l'approche indirecte, en dissociant le travail sur la langue-source et sur la langue-cible, permet d'étendre les systèmes à une pluralité de langues (cf. *infra*, § 3.1.).

1.3. De 1975 à nos jours : la reprise

Depuis 1975, on a pu assister à une reprise assez massive des travaux dans le domaine de la traduction automatique : activités de recherche et développement, commercialisation de produits. Cette reprise s'explique à la fois par une augmentation des **besoins** des entreprises privées et publiques en matière de traduction (et donc par un accroissement du marché et des sources de financement), et par la sophistication croissante des **outils** informatiques, tant techniques (puissance et rapidité des ordinateurs) que théoriques (développement de langages de programmation plus adaptés, et de modélisations plus poussées).

Les systèmes de traduction automatique nouveaux élaborés pendant cette période la plus récente sont, pour la plupart, des systèmes de "deuxième génération", plus perfectionnés que les précédents ; en particulier certains (comme ARIANE) sont capables d'un certain **auto-contrôle** : grâce à des langages informatiques évolués permettant d'appeler récursivement et de façon contrôlée des sous-grammaires, ils peuvent détecter un cheminement erroné, revenir en arrière, entreprendre un autre cheminement, etc.

Citons, à titre d'exemple, les systèmes suivants : ARIANE (GETA, Grenoble : russe-français, anglais-malais, français-allemand, etc.), METAL (Siemens : allemand-anglais, allemand-espagnol), TITUS (Institut du Textile de France : tous les couples de langues entre le français, l'anglais, l'allemand, et l'espagnol).

Le regain d'intérêt pour la traduction automatique s'est accentué depuis une dizaine d'années. Le début des années 80 a été marqué, de façon significative, par le lancement de plusieurs grands projets nationaux (ou internationaux) de traduction par ordinateur : en France le projet national de traduction assistée par ordinateur (PN-TAO, 1983-87) sous les auspices de l'Agence pour l'Informatique, dans la Communauté Européenne le projet EUROTRA (lancé en 1982 et portant sur les langues de tous les pays de la C.E.E.), au Canada le programme national (lancé en 1985) de traduction assistée par

ordinateur au Centre canadien de recherche sur l'informatisation du travail, enfin — et non des moindres — au Japon sous l'égide de l'Agence pour la Science et la Technologie, le vaste projet de traduction et d'interprétation (traduction automatique, systèmes de visualisation multilingues, banques de données terminologiques, synthèse vocale et reconnaissance de la parole pour l'élaboration de "téléphones traducteurs") coordonné au départ par l'Université de Kyôto ; déjà s'annoncent une "troisième" et une "quatrième" génération de systèmes de traduction, qui devraient être implantés sur des ordinateurs de "cinquième génération" !

1.4. Traduction automatique et intelligence artificielle

Parallèlement à la construction des systèmes de deuxième génération conçus spécifiquement pour la traduction automatique, des recherches étaient menées sur la **compréhension** automatique de textes dans la perspective de l'**intelligence artificielle**. L'intérêt des chercheurs dans ce domaine se portait principalement, du point de vue théorique, sur l'articulation entre sémantique, pragmatique et connaissances d'univers : cf. par exemple les travaux de T. Winograd ou de R. Schank (voir *infra* chapitre 8).

Dans cette perspective ont été construits des petits systèmes de compréhension limités à des univers très restreints, et susceptibles de s'intégrer dans des systèmes de traduction automatique. C'est ainsi par exemple que R. Schank a élaboré pour la compréhension une théorie dite de "dépendances conceptuelles", qui proposait un niveau de représentation sémantique postulée universelle, c'est-à-dire indépendante d'une langue naturelle particulière (cf. *supra* chapitre 5) ; une telle représentation était conçue comme jouant le rôle de "langage-intermédiaire" dans un système de traduction, dénommé MARGIE (cf. R. Schank & R. Abelson 1977).

Il faut toutefois reconnaître que le monde de la traduction automatique et celui de l'intelligence artificielle sont restés, jusqu'à une époque récente, assez largement **étrangers** l'un à l'autre : les méthodes et les cadres théoriques n'étaient pas les mêmes (linguistique informatique *vs.* intelligence artificielle), les objectifs non plus (recherche de systèmes opérationnels de traduction en grandeur réelle *vs.* élaboration de maquettes expérimentales de compréhension, où la traduction joue surtout un rôle de test).

Quelques contre-exemples historiques peuvent cependant être cités. Tout d'abord celui de Y. Wilks à l'Université de Stanford (cf. Y. Wilks 1973), qui avait conçu au début des années 70 le projet de construire un système de traduction automatique dans une perspective d'intelligence artificielle, fondé sur un "langage intermédiaire" de représentation mettant en jeu une grammaire sémantique dite de "préférences" et des règles d'inférence (cf. *supra* chapitre 5). Evoquons également deux systèmes de traduction élaborés au début des années 80 à l'Université de Yale, dans la perspective théorique de

R. Schank : le système SAM (dans le domaine des accidents de voiture) qui recourait à la notion de "scenario" ("script"), et le système MOPTRANS qui, de son côté, utilisait la notion de "MOPs" ("memory organization packets") (sur la notion de "script", voir *infra* chapitre 8 § 2.2).

Si elles ont jusqu'à présent suivi des voies différentes, les recherches en traduction automatique d'une part, et en intelligence artificielle de l'autre, sont sans doute en passe de se rapprocher : c'est en tout cas ce que laissent présager les projets de systèmes de "troisième génération", illustrés notamment par le système MU en cours d'élaboration au Japon. Le principe de tels systèmes est d'intégrer des systèmes experts permettant de gérer de vastes bases de connaissances d'univers, et d'effectuer les inférences nécessaires pour établir la signification du texte d'entrée, par-delà son seul sens linguistique (sur le passage du sens à la signification, et le rôle des inférences pour la compréhension, cf. *infra* chapitre 8). Bien entendu, il faudra pouvoir utiliser de telles connaissances à bon escient, c'est-à-dire moins pour simuler une compréhension totale du texte que pour être en mesure d'en donner une bonne traduction — notamment pour aider à la levée de certaines ambiguïtés (ainsi, par exemple, pour comprendre que dans *La fleur est composée des sépales du calice et de la corolle* le groupe prépositionnel *de la corolle* est coordonné à *sépales* et non pas à *calice*, il faut disposer de connaissances dans le domaine de la botanique).

2. Traduction et ordinateur : objectifs et produits

Pris de façon générique, le terme "traduction automatique" recouvre dans les faits une grande variété de systèmes. Nous considérerons tout d'abord la diversité des modes possibles d'intervention respective de l'ordinateur et de l'humain dans ces systèmes, puis les différents types de besoins en traduction auxquels ils tentent de répondre ; nous donnerons ensuite un bref aperçu sur les types de systèmes commercialisés existants.

2.1. Traduction automatique et traductions assistées

L'emploi de l'expression "traduction automatique" marquait bien au départ l'espoir d'arriver à une automatisation complète du processus de traduction : à strictement parler, la traduction **automatique** implique qu'il n'y ait aucune intervention humaine, par opposition à une traduction **assistée par ordinateur**. Or il est clair qu'à l'heure actuelle, aucun système de traduction ne se passe totalement de l'homme : c'est pourquoi la dénomination "traduction assistée par ordinateur" (T.A.O.) a progressivement gagné du terrain sur celle de "traduction automatique". Toutefois, la distinction entre traduction "automatique" et traduction "assistée" n'est pas aussi tranchée qu'il peut y paraître au premier abord.

D'une part, en effet, selon l'importance du travail qui incombe respectivement à l'homme et à la machine, on peut distinguer deux types de traductions assistées (souvent confondues sous le terme "T.A.O.") : la traduction mécanique **assistée par l'humain** ("human aided machine translation"), où la traduction est effectuée par l'ordinateur, avec l'aide de l'homme (cf. plus bas), et la traduction humaine **assistée par l'ordinateur** — dans une acception stricte du terme — ("machine aided human translation"), qui fait jouer un moindre rôle à la machine ; dans un système de ce type (dit aussi "poste de travail du traducteur"), c'est l'humain qui traduit, en faisant appel à l'ordinateur pour l'aider (accès à des bases de données terminologiques, recherche de termes dans des dictionnaires, consultation de corpus bilingues, etc.). Notons au passage que le développement considérable de ce secteur a donné lieu à l'élaboration et à la commercialisation d'un nombre important d'outils d'**aide à la traduction** (en particulier des banques de données terminologiques bi- ou pluri-lingues, comme par exemple EURODICOTAUM de la Commission des Communautés Européennes, ou TERMIUM de Montréal).

D'autre part, les systèmes se différencient selon le **moment** de l'intervention humaine. Certains systèmes font appel à l'humain **pendant** le cours même de la traduction, pour prendre un certain nombre de décisions (désambiguïsation d'un terme, confirmation d'un choix de structure, etc.) : c'est ce qui se passe par exemple avec des systèmes interactifs qui, face à une difficulté de traduction, s'interrompent pour interroger l'humain. C'est à de tels systèmes (qui sont à strictement parler des systèmes de traduction mécanique assistée par l'humain) que l'on réserve parfois restrictivement la dénomination de traduction "assistée", pour les distinguer de ceux qui, au contraire, ne nécessitent d'intervention humaine qu'en **amont** (phase de préparation ou de "pré-édition" du texte d'entrée) et en **aval** (phase de révision ou de "post-édition" du texte de sortie) de la traduction proprement dite, la machine effectuant seule l'intégralité du processus de transformation du texte d'entrée en texte de sortie ; pour qualifier ce deuxième type de systèmes, on recourt alors parfois à la dénomination originelle de traduction "automatique".

Mais là encore des différences existent, selon la nature et la quantité du travail humain nécessaire en amont et en aval : la préparation du texte d'entrée peut se réduire à de simples annotations (par exemple introduction de signes typographiques marquant des fins de syntagmes ou de phrases), mais elle peut aussi aller jusqu'à une complète réécriture du texte à des fins de standardisation ; de même le travail sur le texte de sortie peut consister en une simple post-édition typographique, en une révision légère de la traduction (amélioration des tournures stylistiques par exemple), ou aller jusqu'à une révision beaucoup plus lourde, incluant la correction de contresens. Si l'on ajoute à cela qu'une mise à jour des données linguistiques (dictionnaires et grammaires) peut être effectuée *ad hoc* pour adapter le système au texte — ou au type de texte — à traduire, on aperçoit mieux les limites de la dénomination "traduction automatique".

2.2. Types de besoins en traduction

La part plus ou moins grande prise respectivement par la machine et par l'homme dans la traduction s'appréciera différemment, selon les besoins de l'utilisateur. Les systèmes de traduction "automatique" (avec pré- et / ou post-édition) doivent donc être évalués en fonction des types de **besoins en traduction** qu'ils visent à satisfaire. (Rappelons que les besoins en traduction sont en constante augmentation, et que le coût des traducteurs humains est considérable).

Lorsqu'il s'agit de traductions effectuées à des fins de **veille technologique** (traductions de résumés d'articles scientifiques, de brevets, etc.), l'objectif premier est celui de l'intelligibilité du contenu du texte : peu importe la qualité de la forme, ce qui compte c'est que le système soit capable de donner rapidement, à bas coût et en quantité, des traductions compréhensibles quant au fond, de textes tout venant dans le domaine considéré. Dans cette perspective, même des systèmes n'effectuant pas une analyse très "profonde" du texte d'entrée peuvent être suffisants, à condition de posséder un lexique étendu et bien adapté au domaine : les traductions demandées sont des traductions brutes, utilisables comme telles en usage interne (c'est-à-dire sans révision ultérieure) à des fins de filtrage de l'information. La traduction se fait dans le sens de la "version" (de la langue des documents vers celle du veilleur), ce qui, toutes choses égales d'ailleurs, autorise une qualité moindre.

A côté de ces traductions dites pour le **veilleur**, on distingue les traductions dites pour le **réviseur**. Il s'agit ici d'effectuer des traductions de **textes de référence** (textes administratifs, comptes-rendus, notices d'utilisation, manuels d'entretien, etc.) ; l'exigence est donc plus forte puisqu'elle porte également sur la pertinence des tournures linguistiques et que la traduction se fait dans le sens du "thème" (de la langue du traducteur vers celle du public visé). L'objectif ne peut être atteint que par un système de deuxième génération effectuant une analyse assez fine des phrases, et dans des conditions très précises : spécialisation du domaine, et si possible homogénéité des structures linguistiques du texte de départ (qui, le cas échéant, peut être imposée à son rédacteur sous forme d'une syntaxe contrainte, ou obtenue au prix d'une standardisation lors de la pré-édition) ; sinon, une révision par l'humain du texte de sortie est nécessaire (sous réserve naturellement d'un gain financier par rapport à une traduction humaine).

Enfin, dans le cas de traductions de **textes de prestige** (présentations de sociétés, lettres d'affaires, etc. ... sans aller jusqu'à envisager les textes littéraires!), l'exigence porte, en plus, sur la qualité même de la forme du texte, ce qui engage des questions stylistiques et pragmatico-textuelles que les systèmes de traduction actuels sont encore loin de maîtriser ; en tout état de cause, de telles traductions nécessitent fatalement d'importantes interventions humaines tant sur le texte d'entrée que sur le texte de sortie. Pour satisfaire à

ce type de besoins, on tend actuellement à privilégier les outils de traduction assistée par ordinateur (c'est ce que l'on appelle la traduction pour le **traducteur** — l'humain restant l'acteur principal de la traduction : cf. *supra*, § 2.1.) et aussi à développer des systèmes interactifs de rédaction directement bilingue (on parle alors de traduction pour le **rédacteur** : cf. *infra*, § 5.).

2.3. Les systèmes commercialisés : bref aperçu

Nous ne mentionnerons ici que pour mémoire les **traducteurs électroniques de poche** ("portable translators"), qui contiennent un petit nombre (quelques dizaines de milliers) de mots, d'expressions idiomatiques et de phrases-types de la vie quotidienne, pré-enregistrés dans plusieurs langues ; ces astucieuses petites machines, qui peuvent dépanner le voyageur, et éventuellement impressionner le néophyte (surtout lorsque la phrase de sortie est donnée sous forme orale), ne présentent pas grand intérêt linguistique : ce ne sont que des **dictionnaires** d'expressions et de phrases "à trous" toute faites, qui n'effectuent aucun véritable traitement puisque les équivalences interlangues (accompagnées, le cas échéant, de synonymes ou d'expressions apparentées) y sont codées par avance.

En ce qui concerne les systèmes qui effectuent un véritable travail de traduction, on peut schématiquement regrouper les produits existant sur le marché (et ne nécessitant pas d'intervention humaine pendant la traduction elle-même) en trois grandes classes : les systèmes restreints, les systèmes légers et les systèmes lourds (ou mi-lourds).

Les systèmes **restreints**, robustes et simples, atteignent une notable efficacité (bonne qualité de traduction ne nécessitant pas de révision) au prix d'une double limitation sur le texte d'entrée : limitation linguistique (ils imposent une syntaxe contrainte pré-définie) et limitation du domaine (univers fermé très limité, qui induit des restrictions terminologiques). Une illustration-type en est le système canadien TAUM-METEO, destiné à la traduction des bulletins météorologiques d'anglais en français. A titre d'exemple, voici un bulletin traduit par ce système :

- texte d'entrée : *Forecasts for Ontario issued by environment Canada at 11.30 a.m., Wednesday March 31st 1976 for today and thursday. Cloudy with a chance of showers today and thursday. Low tonight 4. High thursday 10. Outlook for friday...sunny.*

- texte de sortie : *Prévisions pour l'Ontario émises par environnement Canada à 11h30, mercredi 31 mars 1976 pour aujourd'hui et jeudi. Nuageux avec possibilité d'averses aujourd'hui et jeudi. Minimum ce soir 4. Maximum jeudi 10. Aperçu pour vendredi...ensoleillé.*

La normalisation du texte d'entrée se fait parfois de façon interactive avec l'humain, lors même de la rédaction du texte : le système intervient alors comme contrôleur syntaxique, et oblige le rédacteur à modifier les

phrases qu'il n'accepte pas (en particulier lorsque le système détecte des ambiguïtés) ; c'est le cas du système TITUS de l'Institut Textile de France (conçu à l'origine comme système documentaire, et limité aux résumés d'articles sur les techniques du textile). Notons que les textes appelés à être traduits tendent de plus en plus à être rédigés dans une langue "simplifiée" (c'est le cas, par exemple, de toute la documentation de BULL). Bien entendu, tout élargissement du domaine d'application risque de mettre en péril de tels systèmes : c'est ainsi que l'extension du système TAUM au domaine aéronautique a échoué.

Les systèmes **légers** (comme par exemple le système WEIDNER) sont des systèmes simplifiés, implantés sur micro-ordinateurs, et placés sous le contrôle de l'utilisateur : celui-ci a la maîtrise directe des dictionnaires (le système ne comporte au départ qu'un petit noyau de dictionnaire, que l'utilisateur complète lui-même en fonction de ses besoins et de son domaine d'application), et bénéficie de la convivialité des logiciels de traitement de texte sur micro-ordinateur. En revanche, les traductions effectuées ne sauraient être de bonne qualité : la traduction se fait mot à mot, et surtout chaque mot ne comporte qu'une seule traduction dans le dictionnaire. A ce compte, les phénomènes de contextualisation du sens, d'ambiguïté, de polysémie ne peuvent pas être pris en compte : le texte de sortie risque donc de contenir des contre-sens, même pour des langues aux structures de phrase assez semblables (ces systèmes étant, en tout état de cause, peu utilisables pour des couples de langues éloignées).

Les systèmes **lourds** (ou mi-lourds), quant à eux, tournent sur des équipements informatiques nettement plus importants. Le plus connu et le plus répandu de ces systèmes est sans doute SYSTRAN. De tels systèmes n'imposent pas de standardisation *a priori* du texte d'entrée, tout en étant moins rudimentaires que les systèmes légers du point de vue du traitement linguistique : ils construisent une représentation (au moins partielle) des phrases du texte à traduire, et s'efforcent de prévoir dans leurs dictionnaires le sens contextuel des mots polysémiques, afin de limiter les risques de contre-sens ; la partie dictionnaires de ces systèmes est d'ailleurs extrêmement complexe et coûteuse à élaborer. Les performances qualitatives de ces systèmes sont meilleures que celles des systèmes légers, sans être vraiment bonnes pour autant.

Parmi ces systèmes lourds, certains (comme SYSTRAN) conviennent à la **veille** technologique, c'est-à-dire pour effectuer rapidement des traductions brutes de textes tout venant dans des domaines techniques définis. Si l'absence de révision est souvent jugée sans conséquence (la mauvaise qualité de la forme n'empêchant pas la compréhension du contenu), elle peut néanmoins être préjudiciable (les traductions n'étant pas exemptes de risques de contresens dus au peu de "profondeur" de l'analyse du texte dans ces systèmes de traduction pour le veilleur ; pour des exemples, voir *infra*, § 3.1.).

D'autres systèmes lourds, comme par exemple LOGOS, METAL ou B'VITAL-AERO, sont destinés, eux, au **réviseur**. Plus perfectionnés que les systèmes de traduction pour le veilleur, ils traduisent moins vite et le texte de sortie, bien que meilleur, appelle une révision pour devenir vraiment acceptable ; à titre de comparaison, si à l'heure actuelle un système de traduction pour le veilleur traduit en moyenne une dizaine de pages en une heure, un système de traduction pour le réviseur traduit en moyenne une page à l'heure, et nécessite une révision de vingt minutes. Malgré leurs perfectionnements, ces systèmes restent fondés sur des connaissances linguistiques relativement élémentaires, et s'appuient sur les régularités de certains types de textes : ils ne sont adaptés qu'à la traduction de grosses masses de textes homogènes comme des manuels d'utilisation ou de maintenance.

On remarquera que la plupart des systèmes commercialisés qui viennent d'être évoqués ont été conçus à l'origine il y a plusieurs décennies, et ont été rôdés et progressivement améliorés au fil des années, faisant l'objet de nombreuses versions successives (à l'exception de TAUM-METEO qui, depuis 1976, fonctionne 24 heures sur 24 sans qu'aucune modification importante y ait été apportée) : le chemin est long, qui mène de la conception de modèles et de prototypes au développement, puis à la commercialisation de systèmes en grandeur réelle (nombre de systèmes de deuxième génération évolués en sont encore au stade du prototype de laboratoire).

En particulier, pour être opérationnels, les systèmes en grandeur réelle doivent comporter des **dictionnaires** nombreux et de grande taille : pour des applications grand public il faut compter un nombre d'entrées qui oscille, selon les cas, entre plusieurs centaines de milliers et plusieurs millions. De plus, une équipe de développement et de maintenance des "**linguiciels**" (grammaires et dictionnaires) est indispensable. A titre d'exemple, voici la liste des dictionnaires que comporte SYSTRAN (d'après. S. Trabulsi 1988) :

- un dictionnaire des "mots de base" associant à chaque mot des informations de nature morphologique, syntaxique et sémantique ; en cas d'homographie ou de polycatégorie (cf. *supra*, chapitre 3), on trouve autant d'entrées distinctes ;

- un dictionnaire "idiomatique" qui ramène à une seule unité lexicale les expressions traitées comme des blocs (par exemple une expression latine comme *sine qua non*, qui ne devra pas être analysée) ;

- un dictionnaire des "inter-relations fortes", qui "bloque" certaines relations syntaxiques à l'intérieur de groupes nominaux particuliers (par exemple la relation entre l'adjectif et le nom dans *hydraulic brake* ; en effet, contrairement aux règles habituelles de l'anglais, si le groupe nominal se complexifie, comme dans *hydraulic brake valve*, l'adjectif porte toujours sur *brake*, et non sur le nom de tête : la bonne traduction est *soupape de frein hydraulique*, et non pas * *soupape hydraulique de frein*) ;

- un dictionnaire des "groupes nominaux", qui ramène à une unité les mots composés comme *pomme de terre* ;

- un dictionnaire "homographique", qui consigne les exceptions à certaines règles grammaticales générales permettant de distinguer plusieurs homographes (ainsi *note*, dans *prendre note*, ne prend pas d'article, contrairement à la règle générale sur l'objet) ;

- une série de dictionnaires "analytiques", qui spécifient les exceptions des mots aux diverses règles grammaticales générales ;

- un dictionnaire "conditionnel", qui contient des informations syntaxiques et sémantiques sur les mots, permettant de choisir leurs équivalents dans la langue-cible, une fois terminée l'analyse du texte en langue-source (ainsi *grow* a-t-il pour traduction de base *grandir*, mais "*grow* + objet animé" a pour traduction *élever* et "*grow* + objet de type plante" *cultiver*).

Ainsi que cette présentation l'aura montré, la réalisation de systèmes opérationnels en traduction, oblige, comme dans tous les autres domaines du traitement automatique, à de nécessaires compromis entre la recherche de la qualité (fiabilité et justesse de la traduction) et le souci de l'efficacité (rapidité et automatisation maximale de la traduction). Aucun système ne saurait satisfaire simultanément à l'ensemble de ces exigences sur un texte tout venant : ou bien l'on traite automatiquement du texte tout venant et l'on perd nécessairement sur la qualité, ou bien il faut renoncer soit à l'automatisation totale, soit à la généralité des textes traités en restreignant de façon drastique non seulement le domaine d'application mais aussi la structure linguistique des textes à traduire. Le verdict énoncé par Y. Bar-Hillel en 1960 reste pertinent : il faut abandonner l'idée d'une machine à traduire universelle entièrement automatique livrant des traductions de qualité.

3. Les différentes architectures des systèmes de traduction

Les différences dans les architectures des systèmes reflètent la façon dont est conçu le processus de traduction : la traduction entre la langue-source et la langue-cible peut s'effectuer directement ou indirectement, par l'intermédiaire d'une représentation-pivot ou d'un module de transfert — et ceci à un niveau plus ou moins "profond" selon les cas.

3.1. Approche directe et approche indirecte

On distingue classiquement, du point de vue de l'architecture des systèmes de traduction automatique, l'approche dite "directe" et l'approche dite "indirecte".

Les systèmes adoptant l'approche **directe** sont conçus pour ne traduire qu'un couple de langues donné (comme par exemple GAT, ou la toute première version de SYSTRAN, construits pour le couple russe-anglais) ; ils effectuent un travail réduit au minimum nécessaire *ad hoc* pour les deux langues — ils n'ont aucune portée générale — en ne distinguant pas les tâches d'analyse et de génération. L'approche directe était caractéristique des systèmes de première génération des tout débuts, dans lesquels elle s'accompagnait d'une vision dite **locale** du texte : l'unité de traduction était le mot, muni d'informations calculées par un examen de son environnement immédiat sur la chaîne. Une traduction procédant ainsi mot à mot se heurte évidemment à de nombreux problèmes, notamment :

- à l'homographie qui, dans la plupart des cas, ne peut pas être résolue dans un cadre aussi étroit : le groupe *la pêche* doit, selon le contexte syntaxique, être analysé comme un syntagme nominal "article + nom" ou comme un syntagme verbal "pronom objet + verbe" (cf. *supra*, chapitre 3) ;

- à la polysémie, dont la résolution nécessite le recours à des règles contextuelles fines (cf. *supra*, chapitre 5) : *élever un enfant* n'a pas le même sens (et ne se traduit pas de la même façon, par exemple en anglais) que *élever la voix* ou *élever les bras*, ou encore *élever une protestation* ;

- à la question de la structuration et de la délimitation des groupes : *la réponse à la question de cet auditeur* ne se parenthèse pas comme *la tendance à l'ivrognerie de cet individu* (cf. *supra*, chapitre 4) ;

- à la nécessité de restructurer la totalité d'une phrase pour certains passages de la langue-source à la langue-cible : *This bed has been slept in by John* ne se traduit pas en français par * *Ce lit a été dormi dedans par Jean* (tous les étudiants de langue connaissent les désastres d'une traduction mot à mot !).

Manifestement inadéquate, l'approche directe a très vite été abandonnée. Les systèmes, plus évolués et plus généraux, qui adoptent une approche **indirecte** effectuent le passage du texte d'entrée en langue-source au texte de sortie en langue-cible par l'intermédiaire de **représentations** de ces textes : ils séparent donc la tâche d'analyse du texte en langue-source (qui construit une représentation de ce texte) de la tâche de génération du texte en langue-cible (qui part d'une représentation de ce texte) ; de la sorte, chacun des deux modules d'analyse et de génération est réutilisable lorsque le système est étendu à de nouvelles langues-cibles ou sources.

Dans les systèmes de deuxième génération évolués l'approche indirecte s'accompagne par ailleurs d'une vision plus **globale** du texte : les représentations du texte prennent en compte la phrase entière (elles peuvent même, théoriquement, se situer au niveau d'une unité plus vaste : interphrase, paragraphe,...).

On remarquera toutefois que, dans les faits, certains systèmes commercialisés se situent à la frontière de la première et de la seconde génération et

adoptent une approche que l'on pourrait qualifier de **mixte :** ils procèdent à une analyse, mais la conduisent à un niveau si peu profond (syntaxe de constituants) qu'ils n'ont qu'une vision **partielle** de la phrase. Dans de tels systèmes, la traduction se fait au niveau des syntagmes ; d'où des difficultés de délimitation et d'identification des syntagmes, qui ne manquent pas d'entraîner des traductions fautives, comme on peut le voir dans les exemples (attestés) suivants :

- texte d'entrée = *On a réalisé des études* —-> texte de sortie = *Studies were made* ; mais si un groupe prépositionnel est inséré à l'intérieur du groupe verbal, celui-ci n'est plus reconnu et la traduction devient fautive (dans le texte de sortie, la construction n'est plus passive mais active, l'équivalent lexical du verbe est mauvais, et un regroupement erroné est effectué, le N *études* étant rattaché au N *soin* par l'intermédiaire d'un *des* compris comme préposition contractée *de* + *les*, et non comme article indéfini) : texte d'entrée = *On a réalisé avec le plus grand soin des études* —-> texte de sortie = *One realized with the greatest care of the studies* ;

- texte d'entrée = *Des groupes et des départements participent au projet* —-> texte de sortie = *Groups and departments take part in the project* ; mais si l'ordre "syntagme nominal + syntagme verbal" est inversé, le groupe nominal sujet n'est plus identifié, et la traduction atteste un regroupement erroné (*les groupes et les départements* rattaché à *projet*, par le biais d'un *des* compris comme *de* + *les*) : texte d'entrée = *Participent au projet des groupes et des départements* —-> texte de sortie = *Take part in the project of the groups and departments.*

On voit dans ces exemples que les groupes ont été rattachés au plus près à gauche, sans vision globale de la phrase ; seule une analyse syntaxique de la phrase entière, telle que la pratiquent les systèmes de deuxième génération évolués, permet d'éviter de telles erreurs.

3.2. Approche par pivot et approche par transfert

A son tour, l'approche indirecte peut être subdivisée en deux : l'approche dite par "pivot" (ou "interlingua") et l'approche dite par "transfert".

Un système adoptant l'approche par **pivot** se compose de **deux** modules seulement : un module d'analyse, qui produit une représentation du texte d'entrée en langue-source dans un **langage-pivot** postulé indépendant de toute langue (ou, à tout le moins, des deux langues source et cible), et un module de génération, qui construit à partir de cette même représentation un texte de sortie en langue-cible (cf. Figure I).

Le pivot est donc conçu comme la **charnière** entre l'analyse et la génération. Il vise à permettre la construction de représentations suffisamment abstraites pour consigner le contenu du texte, indépendamment des contraintes

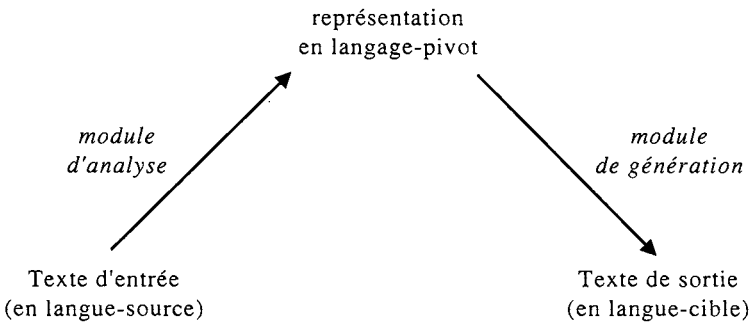


FIGURE I
Architecture d'un système à pivot

d'expression propres à chacune des deux langues source et cible. C'est pourquoi il doit se situer à un niveau conceptuel, et non plus linguistique, et peut, dans son principe, contenir des connaissances non linguistiques concernant le domaine particulier traité.

L'objectif d'une approche par pivot avait été adopté dans les années 70 par certains grands systèmes de deuxième génération, comme par exemple le système russe-français du CETA de Grenoble (1961-1971). Intellectuellement séduisante, la solution du pivot paraissait particulièrement intéressante pour effectuer des traductions entres langues typologiquement éloignées, et économique dans le cas de traductions multi-lingues (le pivot ayant, en théorie du moins, vocation à être indépendant de toutes les langues, il apparaissait comme pouvant être élaboré une fois pour toutes, quelles que soient les langues à traiter). Reposant sur l'hypothèse (théoriquement discutable) de l'existence d'universaux du langage, l'approche par pivot est extrêmement coûteuse car elle nécessite de conduire l'analyse jusqu'à un degré très "profond" — peut-être inaccessible ; aussi pour des raisons pratiques de faisabilité, cet idéal quelque peu mythique a-t-il été assez largement abandonné dans les milieux traditionnels de la traduction automatique (il est révélateur par exemple que le GETA y ait renoncé dès 1971).

Si la recherche d'un langage-pivot a conservé des adeptes, c'est principalement dans les milieux de l'**intelligence artificielle**, pour l'élaboration de maquettes de traduction dans des univers fermés (cf. *supra*, § 1.4.) : on peut considérer qu'il s'agit davantage de "modèles" que de véritables "systèmes". Dans cette perspective théorique, qui s'attache avant tout à la **compréhension** la plus fine et la plus complète possible du texte d'entrée, le recours à une représentation-pivot doit permettre de donner une représentation du texte dans son ensemble, en prenant en compte les phénomènes interphrastiques et discursifs, et de mener l'analyse du texte d'entrée au niveau de "profondeur"

nécessaire pour la compréhension (d'où la possibilité de réutiliser la représentation-pivot dans un autre cadre que celui de la traduction, par exemple dans un système d'interrogation ou de résumé des informations du texte) ; corrélativement, du fait même de cette "profondeur" de la représentation-pivot, la génération du texte de sortie tend à produire des reformulations paraphrastiques du texte d'entrée plus éloignées dans la forme que des équivalents littéraux.

Jusqu'à présent, les concepteurs de systèmes de traduction ont été moins sensibles à ces avantages théoriques qu'aux difficultés d'élaboration de systèmes à pivot ; dépasser le stade de maquettes sur des univers fermés pour construire des systèmes en grandeur réelle applicables à des textes tout venant supposerait en effet d'une part un travail considérable au niveau des deux modules d'analyse et de génération, et d'autre part la mise en oeuvre d'une quantité, actuellement non maîtrisable, de connaissances sur le monde.

Par différence avec un système à pivot, un système adoptant l'approche par **transfert** se compose, lui, de **trois** modules : en plus des deux modules d'analyse du texte en langue-source et de génération du texte en langue-cible, il comporte en effet un module intermédiaire dit de "transfert", qui transforme la représentation R du texte ("sortie" de l'analyse) en une représentation R' du texte ("entrée" de la génération) (cf. Figure II).

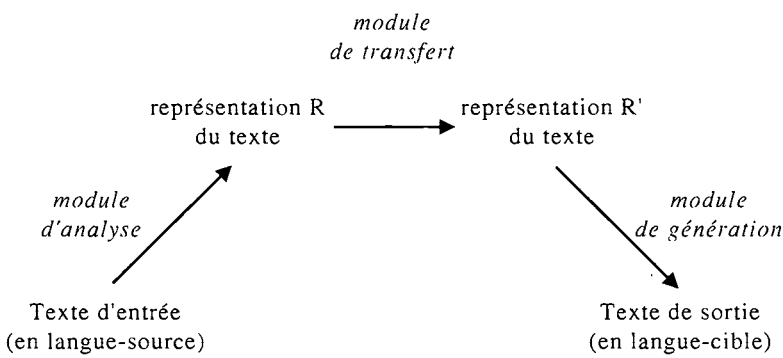


FIGURE II
Architecture d'un système à transfert

Dans leur principe, les systèmes à transfert sont moins ambitieux (ou plus réalistes) que les systèmes à pivot, puisque les deux représentations R et R' sont, *a priori*, distinctes ; elles se situent donc à un niveau moins abstrait que le langage-pivot, et consistent généralement en structures syntaxico-sémantiques de phrases instanciées par des mots de la langue : ce sont des représentations linguistiques plutôt que conceptuelles.

Le module de transfert a pour tâche de transformer la représentation du texte d'entrée en langue-source calculée par l'analyse en une représentation équivalente servant de point de départ à la génération du texte en langue-cible.

Ainsi que nous l'avons dit plus haut, hormis les systèmes conçus pour tester la compréhension, la plupart des grands systèmes de traduction ont opté dès les années 80 pour l'approche par transfert : le système du GETA dès 1971, le système METAL en 1978, le système MU de l'Université de Kyoto dès ses débuts en 1980 ; il en va de même pour le système EUROTRA en 1982.

Toutefois, plus les représentations des systèmes à transfert deviennent "profondes", plus elles tendent à se rapprocher, de fait, d'un pivot : la différence entre les deux types de systèmes ira donc en s'estompant à mesure que les systèmes intègreront des connaissances linguistiques de plus en plus fines et, par-delà ces connaissances linguistiques, des connaissances pragmatiques et d'univers, comme visent à le faire les systèmes dits de "troisième génération".

4. Les traitements effectués par les différents modules

Nous considérerons successivement les traitements effectués par les deux modules d'analyse et de génération, puis les questions spécifiques au processus même de traduction, qui sont traitées dans le module de transfert.

4.1. Les modules d'analyse et de génération

L'analyse et la génération ne sont pas, en soi, spécifiques de la traduction : ces deux tâches peuvent se retrouver dans d'autres types d'applications ; pour un exposé général des problèmes posés par l'élaboration de systèmes d'analyse et de compréhension de textes d'un côté, de génération de textes de l'autre, nous renvoyons respectivement aux chapitres 10 et 11 *infra*. Nous considérerons ici uniquement ce qui fait la spécificité des modules d'analyse et de génération en traduction automatique.

Cette spécificité réside en ceci que l'ensemble des tâches à accomplir est **conditionné par le texte d'entrée**, puisque l'objectif même de ces systèmes est de calculer le contenu de ce texte, pour le traduire. Pour le module d'**analyse**, il en résulte que la représentation qu'il a pour mission de construire doit respecter le contenu du texte d'entrée. Comme on l'a vu plus haut (cf. *supra*, § 2.2.), la question du dosage entre respect du fond et respect de la forme reçoit, de fait, des réponses variables selon les besoins de l'application ; corrélativement, l'élaboration du module d'analyse pour un système de traduction conduit à s'interroger sur le niveau de "profondeur" optimal de la représentation finale du texte d'entrée construite par ce module (nous aurons

l'occasion d'y revenir plus loin). En ce qui concerne le module de **génération**, la représentation du texte qui lui sert d'entrée constitue pour lui une donnée déjà construite ailleurs (c'est-à-dire par le transfert), et induite en définitive par le texte d'entrée : la question du "quoi dire?" ne se pose donc pas pour lui, contrairement à ce qui se passe, par exemple pour des modules de génération intégrés à des systèmes d'interrogation de bases de données (cf. *infra*, chapitre 9). Là encore, l'importance du travail qui revient à ce module dépend du niveau de "profondeur" de la représentation du texte : plus cette représentation est "profonde", plus sa tâche est complexe.

Précisons en outre que certains chercheurs travaillent à l'élaboration de modules d'analyse et de génération **réversibles**, utilisables pour des traductions bi-directionnelles (langue A <—-> langue B) : c'est le cas, par exemple, dans EUROTRA (cf. L. Danlos & O. Laurens 1991 pp.10-11) ; l'idée est de construire des algorithmes symétriques, et de faire partager les mêmes grammaires et les mêmes dictionnaires par le module d'analyse et le module de génération d'une langue donnée (cf. M. Dymetman 1992).

Chacun des deux modules d'analyse et de génération se compose lui-même d'un certain nombre de **sous-modules** : on ne passe pas en une seule étape du texte en langue-source à sa représentation R, ni de la représentation R' au texte en langue-cible ; chaque sous-module construit donc une **représentation intermédiaire** du texte. Sans préjudice du traitement éventuel de connaissances non-linguistiques, les différentes étapes ainsi parcourues (en sens inverse) lors par l'analyse et la génération dans les systèmes évolués réalisent le traitement des **connaissances linguistiques** classiquement répertoriées dans tous les traitements automatiques (cf. *supra* chapitres 3 à 5) : morphologie (identification des mots et calcul des informations qui leur sont associées), syntaxe superficielle (structuration en constituants et calcul des fonctions), syntaxe profonde (c'est-à-dire sémantique des relations syntaxiques : calcul des relations prédicatives, et des valeurs actancielles ou des cas) et sémantique lexicale (calcul du sens des unités lexicales). Ces différents calculs se font grâce aux données linguistiques que constituent les **dictionnaires** et les **grammaires**.

S'il y a un large consensus concernant la nécessité d'une prise en compte de ces divers types de connaissances linguistiques pour l'analyse de la langue-source et la génération de la langue-cible, en revanche les divers systèmes de traduction se différencient par la manière dont ils **structurent** en sous-modules le traitement de ces connaissances : tous ne se donnent pas le même nombre de niveaux intermédiaires, et tous n'articulent pas les sous-modules de la même façon. Pour les uns (par exemple le système EUROTRA), qui adoptent une approche **stratificationnelle**, il s'agit de modules strictement **hiérarchisés**. Pour les autres au contraire, les divers sous-modules **coopèrent** : ainsi dans les systèmes ARIANE, METAL ou LOGOS, tous les niveaux de représentation sont calculés simultanément — certaines

règles paraissent en effet plus faciles à mettre en oeuvre en jouant sur plusieurs niveaux en même temps (résolution d'ambiguïtés syntaxiques grâce à des connaissances d'ordre sémantique, etc. : cf. *supra*, chapitres 3 à 5).

En ce qui concerne la **couverture** des phénomènes linguistiques, les systèmes de traduction automatique ont, théoriquement du moins, vocation à couvrir la totalité de la langue-source et de la langue-cible (même si, en pratique, il arrive que les grammaires et les dictionnaires ne couvrent qu'une partie de la langue, c'est-à-dire un "sous-langage" restreint à certaines structures de phrase et à certaines unités lexicales liées au domaine : cf. *supra*, § 2.3., à propos des systèmes commercialisés). A cet égard, mentionnons l'exemple d'EUROTRA, où est conduite de façon systématique la description linguistique des constructions et des marqueurs des langues traitées (cf. références dans les repères bibliographiques en fin de chapitre).

Du point de vue des **théories linguistiques** enfin, si l'élaboration de systèmes de traduction a, pendant longtemps, été du ressort du "bricolage" empirique (emprunts hétérogènes, pour chaque niveau de description, à diverses écoles ainsi qu'à la grammaire traditionnelle), on observe à l'heure actuelle une tendance à suivre de façon plus stricte et cohérente les travaux linguistiques conduits dans une ligne théorique donnée. On assiste en particulier dans les recherches les plus récentes à un relatif déclin des modèles purement syntaxiques inspirés par la grammaire générative classique (et, corrélativement, des structures arborescentes comme représentations-types des données), au profit d'une part de perspectives lexicalistes (on parle de systèmes de traduction automatique "pilotés par le lexique" : "lexicon-driven machine translation systems") et des sémantiques formelles d'autre part ; c'est ainsi par exemple que le système ROSETTA (développé par Philips aux Pays-Bas) prend explicitement appui sur des grammaires de Montague, tout en y apportant certains aménagements pour les adapter aux besoins d'un système de traduction opérationnel.

4.2. Le module de transfert

Le module de transfert reçoit en entrée la représentation R du texte en langue-source calculée par le module d'analyse, et il la transforme en une représentation R' équivalente qui sert ensuite de point de départ à la génération du texte en langue-cible.

En théorie, un système de traduction indirect peut effectuer le transfert entre deux représentations R et R' du texte à n'importe quel niveau de "profondeur" (cf. M. Nagao 1983 p. 1532) : le niveau le plus "superficiel" serait celui de la simple translitération de formes (par exemple en japonais le passage automatique des caractères *kanji* aux caractères *kana* ou aux caractères romains), le plus "profond" serait celui d'un "pivot" universel codant le contenu indépendamment de toute langue et effectuant un simple transfert lexical ; entre les deux se situent divers niveaux intermédiaires (transfert mot

à mot, transfert syntaxique entre constituants immédiats, transfert syntaxico-sémantique, transfert contextuel).

Plus les représentations R et R' sont "profondes", plus les opérations de transfert ont de chances d'être légères. Comme les modules d'analyse et de génération sont écrits une fois pour toutes pour une langue donnée, alors que pour tout couple de langues particulier il faut élaborer un module de transfert spécifique, le but visé est évidemment la légèreté maximale du transfert (souvenons-nous par exemple que le système EUROTRA, qui travaille sur 9 langues, doit élaborer 9 modules d'analyse, 9 modules et génération et 72 modules de transfert !). Toutefois, comme on l'a vu à propos de la recherche d'un "langage-pivot", il n'est pas non plus souhaitable de complexifier *ad infinitum* l'analyse et la génération : ici encore, il faut donc se résoudre à des compromis. Sans prétendre accorder une valeur absolue à de tels décomptes, les chiffres suivants nous paraissent révélateurs de ce problème : selon S. Trabulsi (1988), la traduction d'un mot nécessiterait en moyenne 25 000 à 30 000 opérations dans un système de première génération, 1,5 million dans un système de deuxième génération et 100 millions d'opérations dans un système de troisième génération !

Le maximum de légèreté pour le transfert consiste à relier deux représentations isomorphes qui ne diffèrent que par le **vocabulaire** ; il n'y a alors qu'à opérer une traduction des unités lexicales : c'est en quelque sorte le "degré zéro" du transfert. Toutes choses égales par ailleurs, cet objectif a des chances d'être plus facilement réalisable sur des langues typologiquement proches. Mais, même entre des langues apparentées, on sait que les constructions peuvent différer sensiblement : le module de transfert doit alors faire jouer également des règles de transformation de **structures**.

Dans les systèmes de deuxième génération évolués (comme par exemple ARIANE ou EUROTRA), qui opèrent le transfert à un niveau syntaxico-sémantique, les représentations (de phrases) sont des représentations de type "prédicat-argument(s)-(modifieur(s))" (cf. *supra*, chapitre 5) : ce sont, très classiquement, des **arborescences** "décorées" dont les noeuds correspondent à des unités lexicales auxquelles sont associées des informations ("attributs") morphologiques, syntaxiques et sémantiques (catégorie, voix, genre, nombre, temps, détermination, rôle actanciel, etc.). Un certain nombre de formes de surface n'apparaissent pas comme unités terminales, à ce niveau : d'une part celles qui fonctionnent comme "valeurs" d'attributs d'une unité (déterminants du nom, flexions du verbe, etc.), d'autre part celles qui dépendent de la construction syntaxique d'une unité (par exemple les prépositions fortement régies). Ainsi dans la représentation de la phrase *Jean a donné un livre à Marie*, n'apparaîtront comme unités terminales que *donner* (prédicat), *Jean* (argument-1), *livre* (argument-2) et *Marie* (argument-3) — une information syntaxique associée à l'entrée lexicale *donner* spécifiant que son argument-3 se construit avec la préposition *à*.

Le choix de ce type de représentations a conduit à privilégier la question du transfert des structures prédicatives. Rappelons, à la suite de L. Danlos (1989, pp.107-110), que l'isomorphisme des représentations, qui rend le transfert **simple**, peut exister même si les constructions syntaxiques superficielles diffèrent (ainsi dans *Jean se souvient de ce jour / John remembers that day*, dans *Mon chien m'obéit / My dog obeys me*, ou dans *Je veux que tu viennes / I want you to come*), ou si une unité simple dans une langue correspond à une expression composée dans l'autre (ainsi *se rendre compte de / realize*). Quant au non-isomorphisme des représentations, qui rend le transfert **complexe**, il peut être illustré par un classique de la traduction entre langues germaniques et langues romanes, à savoir la non-correspondance entre les prédicats de la langue-source et de la langue-cible (comme : *He swam across the river / Il a traversé la rivière à la nage* ; *John hammered the nail into a plank / Jean a enfoncé le clou dans une planche avec un marteau ; John pushed the door open / Jean a ouvert la porte en la poussant* ; *John kissed Mary goodbye / Jean a dit au revoir à Marie en l'embrassant*).

Ces questions de transfert entre structures **prédicatives** ont beaucoup retenu l'attention en traduction automatique ; des études linguistiques fines (comme par exemple celles menées sur les verbes-support dans le cadre d'EUROTRA : cf. L. Danlos, *ibidem*) permettent d'augmenter le nombre de cas de transferts simples, en élaborant des représentations suffisamment "profondes" pour devenir structurellement communes aux deux langues à décrire. Toutefois, les problèmes rencontrés au niveau du transfert ne se réduisent pas, loin s'en faut, aux seules questions d'équivalence entre structures prédicatives. La prédominance, dans les systèmes de deuxième génération, des représentations de type arborescentes (inspirées des modèles syntaxiques), a conduit, de fait, à reléguer au second plan un certain nombre de phénomènes tout aussi importants pour la traduction, notamment les phénomènes énonciativo-référentiels, les phénomènes lexicaux et les phénomènes discursifs.

Les phénomènes **énonciativo-référentiels** concernent, on le sait, l'assignation de valeurs référentielles (de temps, d'aspect, de détermination, de modalité, etc.) à travers l'emploi d'un certain nombre de marqueurs grammaticaux. Tous les enseignants de langue savent combien les langues, même proches, peuvent différer dans l'emploi de tels marqueurs, et combien ceux-ci sont **polysémiques** : les utiliser à bon escient dans une langue étrangère est un signe de maîtrise de cette langue. Or le transfert de ces marqueurs d'une langue à l'autre est un problème mal résolu dans le cadre d'un transfert entre structures prédicatives. Pour le traiter de façon adéquate, on ne saurait en effet se contenter d'étiquettes (traits binaires) du type "+/- perfectif", "+/- déterminé", etc., décorant les noeuds des arbres ; il n'est, pour s'en convaincre, qu'à considérer les quelques exemples suivants de non-correspondance entre marqueurs :

- équivalents anglais du présent français : *Les fabricants de circuits intégrés couvrent 30% de leur marché national / Integrated circuit manufactu-*

rers supply 30% of their own home market (présent simple) ; *A un moment où les Etats-Unis et le Japon prennent de nouvelles initiatives (...), l'Europe ne peut plus se contenter (...) / At a time when US and Japan are taking new initiatives (...)* (présent progressif) ; *En 1983, la France lance un nouveau programme de communication / In 1983 France launched a new communication programme* (prétérit) ; *Ils viennent voir le Ministre / They have come to see the Minister* (present perfect) ; *Cette situation s'accentue depuis plusieurs années / This situation has been developing over a period of years* (present perfect progressif) ;

- équivalents anglais de l'article défini allemand : *Das Wasser in der Tasse ist schmutzig / The water in the cup is dirty* (article défini), mais *Das Wasser im Rhein ist schmutzig / Water in the Rhine is dirty* (pas d'article) ;

- équivalents espagnols de l'article zéro anglais : *Many linguists prefer structural semantics / Muchos lingüistas prefieren la semantica estructural* (article défini), mais *She studied semantics for three semesters / Ella estudió semantica durante tres semestres* (article zéro).

Subtils à décrire, ces phénomènes n'en obéissent pas moins à des régularités : chaque langue possède son système propre, et la mise en correspondance de deux langues ne peut se faire en termes d'étiquetages sémantiques, mais seulement d'**opérations** (linguistico-cognitives) inter-reliées en contexte. Aussi est-ce en direction de systèmes de règles permettant de calculer des équivalences entre de telles opérations que s'orientent actuellement les réflexions des linguistes participant à des recherches en traduction automatique (voir par exemple l'article de F. Lab (1990) sur le transfert des temps français en anglais, et le document de C. Zelinsky-Wibbelt (1991) sur le transfert des déterminants allemands en anglais et en espagnol, auxquels nous avons emprunté les exemples ci-dessus) ; ces travaux font ressortir la nécessité d'intégrer des descriptions fines des phénomènes concernés, en tenant compte de l'apport des théories linguistiques et en particulier des études contrastives.

De la même façon, les phénomènes **lexicaux** d'une part, et **discursifs** de l'autre ne reçoivent sans doute pas le traitement linguistique le plus adéquat dans le cadre des représentations prédicatives classiques : la sémantique lexicale est habituellement traitée en termes de traits binaires et de taxinomies rigides — traitement fort réducteur, en particulier pour rendre compte des phénomènes de **polysémie** lexicale ; les anaphores, reprises interphrastiques, thématisations, ellipses et autres opérations textuelles obligent, de leur côté, à dépasser le cadre strict de la phrase (pour de plus longs développements sur ces questions, nous renvoyons au chapitre 5 *supra* consacré à la sémantique, ainsi qu'aux chapitres 8 et 9 *infra*, où elles sont présentées respectivement dans la perspective de la compréhension et dans celle de la génération).

Pour rendre compte de ces phénomènes linguistiques et, *a fortiori*, pour doter les systèmes de connaissances pragmatiques (qu'il s'agisse de pragmatique linguistique, de connaissances d'univers ou de connaissances encyclo-

pédiques), ainsi que de mécanismes de raisonnement, il faut élaborer de nouveaux types de représentation du texte. Et l'on retrouve ainsi la question du niveau optimal de "profondeur" où doit se situer la représentation du texte, pour effectuer une bonne traduction : jusqu'où faut-il comprendre pour traduire ? et qu'est-ce qu'une bonne traduction ? (pour qui ? pour quoi faire ?). Ainsi qu'en témoigne la diversité des points de vue rassemblés dans Y. Lepage (ed.) (1990), ce débat, qui sous-tend un certain nombre de clivages (transfert / pivot, traduction automatique / compréhension automatique, linguistique informatique / intelligence artificielle), est loin d'être clos.

L'exemple de l'**ambiguïté**, véritable pierre d'achoppement des systèmes de traduction, est à cet égard révélateur. S'il est indispensable de lever toutes les ambiguïtés pour comprendre (au sens plein de ce terme) un texte, en revanche cela n'est pas toujours nécessaire pour le traduire ; certaines ambiguïtés peuvent en effet (si les formes de la langue-cible s'y prêtent) être conservées et transférées telles quelles : ainsi par exemple pour traduire en français *The soldiers fired at the tourists and they fell*, on peut ne pas recourir aux connaissances d'univers et aux inférences nécessaires au calcul correct de l'anaphore *they*, le pronom français *ils* conservant l'ambiguïté (ce qui ne serait pas le cas avec *The soldiers fired at the women and they fell*, où il faudrait choisir entre *elles* et *ils*). Il en va de même pour un certain nombre d'ambiguïtés syntaxiques, dès lors que les deux langues présentent des constructions se prêtant aux mêmes dualités de structuration. La traduction n'exige donc pas dans tous les cas un niveau de "profondeur" comparable à celui de la compréhension.

Cette constatation, bien connue au demeurant des traducteurs humains, nous conduit à une dernière question, que la vogue actuelle des recherches **cognitives** fera peut-être émerger un jour : quels sont les mécanismes psycholinguistiques constitutifs de la traduction humaine, et quelle pertinence y aurait-il à vouloir les simuler au niveau des procédures d'un système de traduction automatique ?

5. Perspectives

Plusieurs voies de recherche nouvelles s'ouvrent actuellement dans le domaine de la traduction automatique. Nous mentionnerons pour mémoire l'arrivée en force, depuis le début des années 90, de systèmes probabilistes et de systèmes dotés de certaines capacités d'apprentissage ; nous évoquerons seulement ici la conception de nouveaux types de systèmes de traduction assistée par ordinateur, et l'élaboration de méthodes d'évaluation des systèmes de traduction.

En matière de **traduction assistée par ordinateur**, s'élaborent à l'heure actuelle, à côté des systèmes d'aide aux traducteurs professionnels, un certain

nombre de systèmes **interactifs** de traduction pour le grand public : on est ici dans le domaine de la "traduction personnelle". Certains de ces systèmes visent à une aide au **rédacteur** (cf. Ch. Boitet 1990) ; l'idée étant que la traduction s'effectue à mesure même que se construit le texte, à travers un **dialogue** entre l'utilisateur et la machine (les systèmes devront alors intéger des modèles de dialogue) — l'ordinateur pouvant imposer certaines constructions, poser des questions face à une difficulté, etc., et le rédacteur pouvant de son côté introduire dans le système des connaissances linguistiques pertinentes pour le traitement du texte, corriger à tout moment la traduction en train de se faire : c'est ce que l'on appelle la **rédaction bilingue automatique** (un projet de ce genre est en cours de réalisation au Canada). Ces systèmes utiliseront des traitements de texte, ou recourront à l'hypertexte ; certains pourraient comporter un module de synthèse vocale utile, par exemple, pour la levée des ambiguïtés qui existent à l'écrit, mais sont levées par la prosodie. De tels systèmes pourraient également être utilisés dans le cadre de l'apprentissage assisté par ordinateur. Enfin, on songe à faire effectuer par les systèmes ce que l'on appelle des **rétrotraductions** : le texte traduit en langue-cible étant retraduit en langue-source, afin de permettre au rédacteur de corriger la traduction (notamment les éventuels contresens) par confrontation du texte initial et du texte final en langue-source.

La seconde piste de recherche concerne l'élaboration de méthodes d'**évaluation** des systèmes de traduction automatique, par des évaluateurs extérieurs, qui n'ont pas accès aux linguiciels, et traitent les systèmes étudiés comme des "boîtes noires". Indépendants des concepteurs des systèmes, ces chercheurs tentent de constituer de façon rigoureuse des **batteries de tests** fondés sur des critères linguistiques et sur des expérimentations systématiques de corpus de traductions (cf. M. King & K. Falkedal 1990).

De tels travaux devraient conduire à dégager les "zones d'ombre" des différents systèmes, en particulier en ce qui concerne la cohérence et la finesse des descriptions linguistiques des langues traitées. En la matière, beaucoup reste à faire, notamment pour introduire dans les systèmes une véritable **sémantique linguistique** dans une perspective **contrastive**. Seules des descriptions minutieuses situées à un niveau "microscopique" peuvent expliquer en termes de régularités sous-jacentes ce qui, considéré à un niveau "macroscopique", n'apparaît encore que comme des "cas particuliers" ou des "exceptions" aux règles de base. A la question "Pourquoi la traduction automatique n'est-elle pas plus répandue (et) pourquoi (...) n'existe-t-il pas davantage de systèmes opérationnels?", la réponse que donnait M. Nagao en 1983 (p.1532), reste entièrement d'actualité : "La raison générale en est notre **méconnaissance des lois du langage**. Trop peu de phénomènes peuvent s'expliquer à l'aide de règles grammaticales de base".

Catherine FUCHS
(ELSAP-CNRS)

Repères bibliographiques

1. Présentations de synthèse sur la traduction automatique :

[Les articles que nous avons sélectionnés ci-dessous sont d'une lecture très accessible pour les non-spécialistes.]

BOITET, Ch. (1992) : Traduction : les machines ne font pas dans la nuance, *Sciences et Avenir*, 86, Paris, 66-71.

NAGAO, M. (1983) : La traduction automatique, *La Recherche*, 150 : 14, Paris, 1530-1541.

SLOCUM, J. (1986) : Is machine translation linguistics ?, *Computational Linguistics*, 12 : 2, 125.

TRABULSI, S. (1988) : Techniques de la traduction automatique, *AFCET / Interfaces*, 66, 6-12.

[L'ouvrage suivant est plus spécialisé :]

KING, M. (ed.) (1987) : *Machine translation today : the state of the art*, Edinburg, Edinburg University Press.

2. Histoire de la traduction automatique :

[Outre les titres qui suivent, signalons l'existence d'une revue internationale consacrée à la traduction automatique : *Machine Translation*, publiée depuis 1962 par l'Association for Machine Translation and Computational Linguistics.]

ALPAC (1966) : *Language and machines ; computers in translation and linguistics*, Report by the Automatic Language Processing Advisory Committee Division of Behavioral Sciences, Washington, National Academy of Sciences, National Research Council

[Le célèbre rapport américain qui marqua un coup d'arrêt dans les recherches en traduction automatique.]

BAR-HILLEL, Y. (1960) : Some reflections on the present outlook for high quality machine translation, *Advances in computers*, 1, New-York, Academic Press, 91-163.

[Premier cri d'alarme mettant en question la possibilité de traductions entièrement automatiques de bonne qualité.]

BOITET, Ch. (ed.) (1989) : *Bernard Vauquois et la T.A.O. : 25 ans de traduction automatique; Analectes*, Grenoble, Université de Grenoble I & Association Champollion.

[Recueil des principaux articles de B. Vauquois, qui dirigea le GETA de Grenoble jusqu'à sa mort en 1985 ; permet en particulier de suivre l'évo-

lution dans la conception des systèmes de ce centre, et de les situer dans le contexte international.]

KING, M. (1989 / 1990) : Traduction assistée par ordinateur, *Tribune des Industries de la Langue*, 1 : 1 / 3, Paris, Observatoire Français des Industries de la Langue, 2-4.

[Courte note sur la situation de la traduction par ordinateur dans la seconde moitié des années 80.]

MACKLOVITCH, E. & ISABELLE, P. (1990) : Les voies actuelles de la traduction automatique du Canada, *Tribune des Industries de la Langue*, 2 : 4 / 7, Paris, Observatoire Français des Industries de la Langue, 22-25.

[Bref historique des recherches en traduction automatique au Canada.]

3. Traduction automatique et traitement de la langue :

- Traduction automatique et intelligence artificielle :

LEPAGE Y. (1990) : La sémantique dans les systèmes de traduction automatique relevant de l'approche "deuxième génération" et de l'approche "intelligence artificielle", *T.A. Informations*, 31 : 1, Paris, Klincksieck, 39-48.

SCHANK, R. & ABELSON, R. (1977) : *Scripts, plans, goals and understanding*, Hillsdale, Lawrence Erlbaum.

WILKS, Y. (1973) : An artificial intelligence approach to machine translation, dans Schank, R. & Colby, K. (eds.) : *Computer models of thought and language*, San Francisco, Freeman.

- Traduction automatique et sémantique :

BOURQUIN, G. (1990) : Du transfert à la traduction : quelle(s) sémantique(s) ?, *T.A. Informations*, 31 : 1, Paris, Klincksieck, 57-69.

DANLOS, L. (1990) : Degré d'abstraction des représentations intermédiaires en traduction automatique ; un exemple : Eurotra, *T.A. Informations*, 31 : 1, Paris, Klincksieck, 25-37.

LAB, F. (1990) : Le temps de la linguistque, *T.A. Informations*, 31 : 1, Paris, Klincksieck, 49-55.

LEPAGE, Y. (ed.) (1990) : "Sémantique et traduction automatique", *T.A. Informations*, Paris, Klincksieck, 31 : 1.

STEINER *& al.* (eds.) (1988) : *From syntax to semantics ; insights from machine translation*, Londres, Pinter.

ZELINSKY-WIBBELT, C. (1991) : Reference as a universal cognitive process : a contrastive study of article use, *Eurotra-Deutschland working papers*, Saarbrücken, I.A.I., 21.

- Réversibilité des modules d'analyse et de génération :

DYMETMAN, M. (1992) : *Transformation de grammaires logiques et réversibilité en traduction automatique*, Thèse d'Etat, Université de Grenoble.

- Evaluation des systèmes de traduction automatique :

KING, M. & FALKEDAL, K. (1990) : Using test suites in evaluation of machine translation systems, *Actes du Congrès COLING*, vol. 2, Helsinki, Association for Computational Linguistics, 211-216.

4. Présentations de systèmes de traduction automatique :

[Les publications indiquées ci-dessous, du fait même de leur spécialisation, ne sont pas toujours très accessibles pour un lecteur non-spécialiste.]

LAWSON, V. (ed.) (1982) : *Practical experience of machine translation*, North-Holland.

[Recueil déjà un peu ancien, mais intéressant par les exemples qu'il donne de traductions effectives produites par des systèmes.]

- Système ARIANE :

[Système de deuxième génération évolué conçu en 1978 et développé au GETA de Grenoble ; la version la plus récente est Ariane-G5 ; une application tourne à l'Aérospatiale. Outre le recueil d'articles de B. Vauquois (Ch. Boitet (ed.) 1989), et les articles cités plus haut de S. Trabulsi (1988 pp. 9-10) et de M. Nagao (1983 pp. 1536-1537), le lecteur pourra se reporter à :]

GUILBAUD, J-Ph. (1986) : Le modèle allemand-français du GETA comme exemple de système de traduction automatique de deuxième génération, dans C. Fuchs (ed.) : *Informatique et recherche textuelle*, Caen, Centre de Publications de l'Université, 31-67.

[Les résultats de recherche du GETA ont été commercialisés pendant plusieurs années par la société B'VITAL, qui a elle-même été rachetée en 1990 par la société européenne de traduction SITE]

- Système EUROTRA :

[Projet de traduction automatique (deuxième génération évolué) lancé en 1982 par la Communauté Economique Européenne ; concerne les neuf langues de la C.E.E. ; chaque équipe nationale est chargée de l'analyse et de la génération de sa langue, ainsi que du transfert des autres langues vers sa langue.]

DANLOS, L. (1989) : La traduction automatique, *Annales des Télécommunications*, 44 : 1 / 2, 101-110.

JOHNSON, R. *& al.* (1985) : EUROTRA, a multilingual system under development, *Computational Linguistics*, 11 : 2 / 3, 155-169.

KING, M. (1982) : EUROTRA : an attempt to achieve multilingual machine translation, dans V. Lawson (ed.) : *Practical experience of machine translation*, North-Holland.

[Les travaux linguistiques d'Eurotra-France sur le français sont décrits dans les documents suivants :]

DANLOS, L. & LAURENS, O. (1991) : Présentation du projet EUROTRA et des grammaires d'Eurotra-France, *Rapports techniques d'Eurotra-France*, Paris, TALANA, 1.

JORGENSEN, H. (1991) : Les propositions subordonnées circonstancielles : description linguistique et implémentation dans Eurotra, *Rapports techniques d'Eurotra-France*, Paris, TALANA, 2.

LAURENS, O. (1991) : Les phrases comparatives en français ; description linguistique et implémentation dans Eurotra, *Rapports techniques d'Eurotra-France*, Paris, TALANA, 3.

- Système LOGOS :

[Système commercialisé mi-lourd de conception américaine, développé en Allemagne (pour ce qui concerne l'Europe) ; fonctionne sur mini-ordinateurs dédiés ; effectue des traductions destinées à être révisées ; parmi les couples de langues concernés : allemand-français et anglais-français. Vient d'être retenu par le gouvernement canadien et par la Banque d'Allemagne.]

HEARN, Ch. (1987) : Présentation de LOGOS, *Encrages*, 17, Université Paris VIII, 131-141.

- Système METAL :

[Originellement conçu à l'Université du Texas, ce système de deuxième génération, actuellement financé par le groupe Siemens en Belgique, effectue des traductions, destinées à être révisées, dans le domaine des télécommunications et de l'informatique ; concerne surtout les couples néerlandais-français, allemand-anglais et anglais-français]

LAMIROY, B. & GEBRUERS, R. (1989) : Syntax and machine translation : the METAL project, *Linguisticae Investigationes*, XIII : 2, Amsterdam, Benjamins, 307-332.

SLOCUM, J. (1984) : METAL, the LR machine translation system, dans M. King

(ed.) : *Machine translation today : the state of the art*, Edinburg, Edinburg University Press, 319-350.

- Système MU :

[Projet national japonais, lancé en 1982, financé par l'Agence pour la Science et la Technologie, et coordonné au départ par l'Université de Kyôto ; vise la traduction automatique (sans pré- ni post-édition) de textes scientifiques et technologiques du japonais vers l'anglais et inversement ; préfigurerait les systèmes de "troisième génaration" par l'introduction de connaissances pragmatiques et de techniques d'intelligence artificielle dans une version MU-II démarrée en 1990.]

[voir les articles de M. Nagao (1983 p. 1540-1541), de S. Trabulsi (1988 p. 10 et p 12) ; voir aussi les pp. 72-73 de l'article suivant :]

VILLARD, M. (1989) : Traduction automatique et recherche cognitive, *Histoire, Epistémologie, Langage*, 11 : 1, Paris, 56-84.

- Système SYSTRAN :

[Système américain à l'origine, a été utilisé par plusieurs organismes officiels comme l'US Air Force, puis la C.E.E. ; propose en France aux utilisateurs une télé-distribution *via* un micro-ordinateur et un réseau téléphonique ; effectue des traductions brutes à des fins de veille technologique ; concerne une dizaine de couples de langues, dont anglais-français et français-anglais]

[Pour une présentation, voir les articles de S. Trabulsi (1988 pp. 7-9 et p. 11) et de M. Nagao (1983 p. 1534) cités plus haut.]

- Système TAUM-METEO :

[Système de traduction restreint (domaine météorologique entre l'anglais et le français) construit à la fin des année 70 à l'Université de Montréal pour le gouvernement canadien ; les traductions sortant de l'ordinateur ne nécessitent aucune révision.]

CHANDIOUX, J. (1976) : METEO : a translation system of public weather forecast, *FBIS seminar on machine translation*, vol. 1.

[Outre la référence ci-dessus, on peut également consulter l'article de E. Macklovitch & P. Isabelle (1990) cité plus haut.]

- Système TITUS :

[Système conçu et développé par l'Institut Textile de France, à des fins de documentation et de traduction ; restreint au domaine des techniques du textile, il fonctionne de façon interactive avec le rédacteur et impose

une syntaxe contrainte ; concerne tous les couples de langue entre français, anglais, allemand et espagnol.]

DUCROT, J-M. (1982) : TITUS IV, dans Talor, P. & Cronin, P. (ed.) : *Information research in Europ*, Londres, ASLIB.

- Système WEIDNER :

[Américain de conception, racheté par le japonais Bravice, ce système léger de traduction personnelle concerne de nombreux couples de langues, dont français-anglais et anglais-français ; effectue des traductions brutes d'assez mauvaise qualité, nécessitant absolument une révision.]

[voir l'article de S. Trabulsi (1988 p.11).]

- Systèmes de traduction assistée par ordinateur :

ABBOU, A. (ed.) (1989) : "Traduction assistée par ordinateur ; perspectives technologiques, industrielles et économiques envisageables à l'horizon 1990 : l'offre, la demande, les marchés et les évolutions en cours", *Actes du séminaire international*, Paris, DAICADIF.

[Recueil général de présentation, centré sur les enjeux industriels.]

BOITET, Ch. (1990) : Vers la T.A.O. personnelle : le projet LIDIA du GETA, *Tribune des Industries de la Langue*, 3 : 8 / 10, Paris, Observatoire Français des Industries de la Langue, 11-18.

[Court article de présentation d'un projet de système de traduction personnelle pour le rédacteur, lancé au GETA.]

8
COMPRÉHENSION AUTOMATIQUE DE TEXTES

L'analyse sémantique de la phrase isolée traitée hors contexte ne conduit à représenter, on l'a vu au chapitre 5, que la partie de la signification qui peut être directement calculée à partir des seules formes linguistiques (unités et relations) contenues à l'intérieur de la phrase : elle n'épuise donc pas ce que l'on peut appeler la **signification complète** d'un texte, telle que l'humain l'appréhende lors d'un processus de compréhension. Comprendre un texte, c'est en effet, par-delà le simple décodage du contenu littéral de ce qui est dit phrase après phrase, être capable de **relier** les phrases entre elles de façon à reconstruire un tout signifiant et cohérent, et être capable d'**interpréter** le message reçu par rapport à la situation et aux conditions d'énonciation.

Prenons un exemple, emprunté à P. Strawson (1970) : soit la phrase *Le président a exprimé l'opinion que 50 ans est l'âge idéal pour ce poste.* Au niveau du sens littéral, la phrase est parfaitement intelligible : on pourrait la traduire dans une autre langue sans difficulté, sans rien connaître du président ni du poste. Toutefois on accède à une compréhension plus complète si l'on sait d'une part de quel président et de quel poste il s'agit (calcul des références), et d'autre part que le président a précisément un candidat favori âgé de 50 ans (restitution des non-dits : sans doute est-ce là la raison implicite de sa déclaration). Il arrive parfois que la signification non littérale calculée à ce niveau — souvent qualifié de niveau **pragmatique** — déplace, voire même contredise, le sens littéral : c'est ce qui se passe lorsque, selon l'expression courante, on "lit entre les lignes", et que l'on est amené à reconstruire derrière le dit apparent une intention de signification opposée (antiphrase, litote, etc.).

C'est cette signification complète qu'un système de compréhension automatique de textes se donne pour objectif de consigner, sous la forme d'une "représentation interne" du texte à analyser. L'élaboration de systèmes de compréhension automatique de textes écrits (qui suppose par ailleurs effectuées, d'une manière ou d'une autre, les analyses morphologique, syntaxique et sémantique, telles que nous les avons présentées dans les chapitres 3 à 5) se heurte donc, de façon spécifique, à deux types de problèmes que nous n'avons pas encore abordés :

- d'une part, les problèmes liés aux relations **inter-phrastiques** : un texte se construit au fur et à mesure, et chaque nouvelle phrase s'appuie sur les précédentes ;

- d'autre part les problèmes liés au **contexte** : même dans le cas-limite où le texte se réduirait à une simple phrase, sa compréhension en tant qu'énoncé n'en impliquerait pas moins la prise en compte de la situation d'énonciation (identification du locuteur et de l'interlocuteur visé, du lieu et du moment de production du texte), et des conditions d'énonciation (objectif poursuivi par le locuteur, environnement énonciatif au sens large, etc.).

Ces deux types de problèmes sont étroitement liés : les premières phrases d'un texte, une fois énoncées, créent leur propre univers de référence, qui s'ajoute au contexte initial et interagissent avec lui pour produire un nouveau contexte (elles deviennent alors "co-texte") pour la phrase suivante.

Afin d'illustrer de façon concrète ces problèmes, considérons le petit texte suivant (déclaration d'accident automobile) : *Je roulais sur la partie droite de la chaussée quand un véhicule arrivant en face dans le virage a été complètement déporté. Serrant à droite au maximum, je n'ai pu éviter la voiture qui arrivait à grande vitesse.*

(Ce texte est tiré d'un corpus de déclarations d'accident qui a été réuni par les soins de l'équipe "Sémantique des langues naturelles" (des Programmes de Recherche Coordonnée "Intelligence Artificielle" et "Communication Homme-Machine") et qui y fait l'objet d'études pluri-disciplinaires en vue d'un traitement automatique).

Comprendre la signification de ce texte, c'est être capable de se représenter exactement la séquence des événements qui ont conduit à l'accident. Un des tests possibles d'une bonne compréhension est la capacité à **répondre à des questions**, comme par exemple ici :

* Question 1 : "De combien de véhicules est-il question dans ce récit ?" ; Réponse : "Deux".

Cette réponse repose sur une série complexe de calculs, en effet :

- le premier véhicule (appelons-le A) n'est pas nommé explicitement, son existence doit donc être inférée à partir de *je roulais sur la partie droite de la chaussée* (la connaissance des conditions d'énonciation permet de comprendre *rouler* comme renvoyant ici au déplacement d'un véhicule automo-

bile et non, par exemple, d'un tonneau, d'une personne en patins à roulettes, ou d'un homme ivre, et de conclure que le sujet *je* est le conducteur de ce véhicule),

- le second véhicule (appelons-le B) est, lui, nommé deux fois, et de façon différente : d'abord comme *un véhicule* (on comprend qu'il ne s'agit pas de A du fait de la suite *arrivant en face*), puis comme *la voiture* (on comprend qu'il s'agit du même véhicule B en interprétant cette anaphore comme une reprise du genre *véhicule* par l'espèce typique *voiture*).

* Question 2 : "A et B roulaient-ils dans le même sens ?" ; Réponse : "Non, en sens contraire".

Cette réponse suppose que l'on interprète *arrivant en face* comme signifiant "roulant en direction de moi sur la partie symétrique de la chaussée, par rapport au milieu" (et non, par exemple, comme "en face sur le trottoir").

* Question 3 : "Pourquoi B a-t-il été déporté ?" ; Réponse : "Parce qu'il avait pris le virage trop vite".

Cette réponse repose sur la mise en relation de deux zones différentes du texte : d'une part *(arrivant en face) dans le virage* et d'autre part *(qui arrivait) à grande vitesse*, d'où l'on implique une perte de contrôle du véhicule B.

* Question 4 : "Pourquoi A a-t-il serré à droite ?" ; Réponse : "Afin d'essayer d'éviter B qui, étant déporté, arrivait sur la voie où roulait A".

Cette réponse suppose notamment les calculs suivants :

- dans *je roulais sur la partie droite de la chaussée* on comprend qu'il s'agit de "droite" par opposition à "gauche" et non à "courbe" (ce qui aurait été le cas dans *la portion droite de la route*) : si A roule sur la file de droite, c'est qu'il n'y a pas de voiture à sa propre droite, il peut donc "serrer" de ce côté-là,

- on restitue un non dit : B a été déporté vers cette partie de la chaussée où A roulait (et non vers le bas-côté de B),

- *complètement* permet de comprendre que B a franchi le milieu de la chaussée, et s'est retrouvé sur la voie de A,

- enfin *serrant à droite au maximum* se comprend, non pas au sens de "serrer un enfant dans ses bras", mais au sens de la conduite automobile : A s'approche autant qu'il le peut du bas-côté.

* Question 5 : "Dans quel sens était le virage ?" ; Réponse : "Vers la gauche, pour A".

Cette réponse repose d'une part sur toute la représentation spatiale qui s'est trouvée construite peu à peu au fil de la lecture et d'autre part sur la connaissance de lois physiques élémentaires : si B est déporté vers sa gauche (puisqu'il franchit le milieu de la chaussée), c'est que le virage tournait vers sa droite, donc vers la gauche de A.

* Question 6 : "Y a-t-il eu choc ?" ; Réponse : "Oui".

Cela n'est pas dit littéralement, mais se déduit de *je n'ai pu éviter la voiture qui arrivait à grande vitesse* : dans le contexte automobile, *éviter* se comprend comme "ne pas heurter" (et non pas comme "fuir (quelqu'un, ou une discussion)", *je n'ai pas pu* se comprend comme "j'ai voulu, mais n'ai pas réussi à", donc "ne pas réussir à ne pas heurter" équivaut à "heurter effectivement".

* Question 7 : "Le choc a-t-il eu lieu dans le virage ?". Cette question est indécidable, en effet la portée du syntagme prépositionnel *dans le virage* n'est pas claire : doit-on comprendre "en face dans le virage", "arrivant dans le virage", "arrivant en face dans le virage" ou encore "a été déporté (pendant qu'il était) dans le virage" ?

* Question 8 (que nous laisserons à la sagacité du lecteur, et de la compagnie d'assurance) : "Qui est responsable de l'accident ?".

On voit d'après cet exemple que la compréhension du texte s'appuie tout à la fois sur la sémantique des phrases et des liens inter-phrastiques, sur le calcul des références, sur des mécanismes logico-langagiers très généraux ainsi que sur des connaissances non-linguistiques, les unes particulières à un domaine (ou "univers") donné (ici lois physiques et monde de la conduite automobile), les autres générales (connaissances d'ordre encyclopédique) ; toutes ces connaissances aident à "filtrer" la bonne signification en contexte parmi la diversité des significations potentielles, et à **inférer** certaines conclusions à partir de ce qui est dit.

Comme on s'en doute, les traitements informatiques dépendent fortement du type d'**application** visé. En dehors de systèmes construits à des fins purement expérimentales (pour tester les difficultés de la langue et / ou simuler certains aspects de la compréhension humaine), la plupart des systèmes de compréhension de texte ont vocation à être intégrés dans un dispositif plus large. Selon qu'il s'agit d'une interface avec une base de données dans un domaine précis (comme par exemple les horaires de trains) ou d'un programme capable de résumer ou de traduire un texte, on utilisera des stratégies informatiques différentes : dans le premier cas, on s'appuiera sur la connaissance que l'on a de l'objectif du locuteur, ce qui n'est généralement pas possible dans le second cas. Quoiqu'il en soit de cette diversité, le principe général de tous les systèmes reste le même : il s'agit de construire une **représentation interne du texte** (ou du dialogue), plus ou moins déterminée par la tâche à accomplir, en s'appuyant à la fois sur des règles linguistiques et sur des connaissances extra-linguistiques préalablement "entrées" dans la machine.

Pour brosser un panorama synthétique de cette problématique de la compréhension automatique, nous commencerons par rappeler les principaux phénomènes langagiers (pragmatico-textuels) constitutifs de ce que l'on peut

appeler la signification complète du texte (§ 1.), puis nous présenterons les outils de représentation qui ont été élaborés pour en rendre compte (§ 2.), enfin nous proposerons quelques exemples de stratégies de compréhension mises en place dans divers types d'applications (§ 3.).

1. Les phénomènes pragmatico-textuels

Nous considérerons successivement le calcul des relations inter-phrastiques, celui des références, et enfin celui des significations implicites.

1.1. Le calcul des relations inter-phrastiques

Comme nous l'avons dit, tout ce qui précède dans le texte ou le dialogue peut devenir source de référence pour la suite. Ce phénomène existe déjà à l'intérieur de la phrase, en particulier pour les phrases complexes, mais il se déploie plus massivement à l'échelle du texte. Il existe des marqueurs spécifiquement dévolus à ce rôle de reprise **anaphorique**, comme certains pronoms, mais, comme on va le voir, les phénomènes d'anaphore ne se réduisent pas à la seule **co-référence**, et par ailleurs bien d'autres formes peuvent marquer une reprise anaphorique. Rappelons également qu'il existe aussi des phénomènes de **cataphore**, où le pronom ne reprend pas un élément du co-texte antérieur, mais annonce un élément du co-texte ultérieur (ex : *Quand il est entré, Jean avait l'air content*).

Les cas les plus faciles à traiter sont ceux d'anaphore nominale avec un marqueur **grammatical** assurant une **co-référence** (c'est-à-dire où le marqueur de reprise renvoie au même individu que l'élément repris). Rappelons que le marqueur de reprise peut être un pronom personnel (ex : *Un homme est entré. Il était jeune et beau*), ou démonstratif (ex : *Un homme est entré. Celui-ci était jeune et beau*), un article défini (ex : *Un homme est entré. L'homme était jeune et beau*), un adjectif démonstratif (ex : *Un homme est entré. Cet homme était jeune et beau*), etc. Dans de tels cas, les difficultés résident essentiellement dans l'identification de l'élément repris, car le co-texte antérieur présente souvent une pluralité d'antécédents possibles pour le marqueur anaphorique (ex : *Jean a rencontré Paul. Il lui a dit que...*).

Un cas déjà plus difficile à traiter est celui d'une co-référence réalisée par des moyens **lexicaux** (reprise de l'antécédent à l'aide d'un synonyme, d'un hyperonyme, d'un hyponyme, d'une périphrase, etc. (ex : *Marie a laissé son chat pendant les vacances. Le félin / l'animal / le siamois / l'inepte créature a dépéri.*).

Plus difficiles encore sont les cas dits d'**anaphore associative**, où l'antécédent n'est pas explicitement présent dans le co-texte antérieur, mais doit

être reconstruit à partir de ce qui est dit (ex : *Il entra dans un village. L'église était romane* : on comprend "l'église du village" à partir de connaissances d'univers de type "dans un village il y a une église"). Ce phénomène, omniprésent dans les textes, obéit à des règles fines (comparez : *Monsieur Dupont a été tué. L'assassin est en fuite* et *Monsieur Dupont est mort. L'assassin est en fuite*).

Très difficiles à traiter également sont les cas où la reprise anaphorique ne conduit pas à une co-référence. Le marqueur de reprise peut en effet ne pas renvoyer au même individu que l'antécédent : il peut renvoyer à un sous-ensemble de l'ensemble (ex : *Paul a acheté trois stylos. Le premier est déjà cassé*), à un autre objet de la même classe (ex : *Paul avait perdu son stylo. Il en a racheté un*), à un objet du même type (ex : *Le marchand a plusieurs machines à laver en vitrine. Il vend beaucoup la plus chère*), ou simplement à la même notion (ex : *Cet enfant est émerveillé par la neige. Il ne l'avait encore jamais vue*).

Par ailleurs, l'identification des relations anaphoriques concerne également les anaphores autres que nominales : anaphores **verbales** (ex : *Jean a marché sur les mains, et Pierre l'a fait aussi*), anaphores **propositionnelles** (ex : *Jean pense que nous gagnerons, et je le pense aussi*), anaphores **spatio-temporelles** (ex : *ici, là-bas, à deux pas de là, la veille, le lendemain,...*). D'une certaine manière, le jeu des temps verbaux peut être aussi en partie associé aux phénomènes d'anaphores temporelles : en effet, il contribue à situer les procès les uns par rapport aux autres, chaque phrase pouvant éventuellement constituer une nouvelle référence. Au moins trois "positions" spatio-temporelles sont à distinguer pour chaque procès dans un texte (cf. H. Reichenbach 1947) : le cadre de l'énonciation (le ici-maintenant du locuteur), le moment et le lieu où se situe le procès, et le cadre dit de "référence" qui représente le point de vue sous l'angle duquel dans lequel on observe ce procès (ainsi, dans *Je vais à Londres lundi prochain : j'aurai alors fini ce travail*, c'est *Londres* et *lundi prochain* qui constituent le cadre spatio-temporel de référence pour le procès *avoir fini ce travail*). En règle générale, un texte introduit constamment de nouveaux cadres de référence, dans un jeu permanent de changement de point de vue qui rend "vivante" l'évocation des événements décrits : en ce sens, l'anaphore est omniprésente, puisqu'elle est le principal mécanisme qui permet de réaliser ce jeu.

Pour les co-références assurées par des anaphores grammaticales, le traitement de base consiste à établir à partir du contexte antérieur une liste d'antécédents potentiels, à imposer des contraintes sur l'anaphore et à vérifier qu'elles sont satisfaites dans le cas particulier ; les contraintes sont d'ordre morpho-syntaxique et / ou sémantico-pragmatique et peuvent nécessiter le recours à des connaissances d'univers. Ainsi, dans *Il posa l'assiette sur la table et la cassa*, on a deux antécédents possibles pour *la*, du point de vue des contraintes d'accord morpho-syntaxique en genre et nombre, et ce sont nos

connaissances sur la résistance respective des matériaux "assiette" et "table" qui nous permettent de choisir. Plus subtil est l'exemple suivant : *Essuyez bien les poissons, coupez les têtes* (= *des poissons* : anaphore associative), *huilez-les* (= *les poissons sans têtes* : inférence à partir de connaissances sur ce que l'on mange et ce que l'on jette, dans un univers culturel donné) *très légèrement, mettez-les sur le grill.* Lorsque plusieurs candidats restent possibles, certains systèmes imposent de choisir le plus près, d'autres tentent de faire jouer des règles de cohérence par rapport au **thème** de la proposition (sur la prise en compte des phénomènes de thématisation dans les traitements automatiques, voir G. Sabah 1989 chapitre 8 § 1.).

Le calcul des relations inter-phrastiques se heurte au phénomène de l'**ellipse** (absence d'un élément ou d'un membre de phrase, qu'il faut essayer de restituer à partir du contexte). Comme on l'a vu au chapitre 4, certaines ellipses sont "traitables" dans le cadre de la phrase, mais c'est loin d'être le seul cas de figure. En particulier les dialogues sont un domaine privilégié dans lequel les ellipses inter-phrastiques abondent : leur traitement implique alors, comme dans le cas de l'anaphore, la conservation d'éléments du contexte antérieur. Ce traitement est facilité quand on peut s'appuyer sur une connaissance de la tâche. Ainsi dans le dialogue homme-machine avec une base de données, le fait de connaître le type de l'information recherchée permet de restituer plus facilement les éléments de question manquants. Par exemple, dans le dialogue suivant :

— *Quand part le premier train de Paris pour Lyon le 9 octobre ?*
— *à 6h. 30*
— *plutôt après 8h ?*
— *à 8h. 25*
— *et pour Marseille ?*

le fait de savoir qu'une demande d'horaire de train n'a de sens que si l'on fournit une ville de départ et d'arrivée, ainsi qu'une date (et éventuellement) une heure de départ, permet de restituer facilement les questions complètes, simplement en conservant pour chacune de ces informations manquante, la donnée de la question précédente.

Bien entendu ce n'est pas toujours aussi simple, et la résolution des ellipses ne peut se faire que dans le cadre de véritables **modélisations** du dialogue : celles-ci seront abordées dans le chapitre 10.

1.2. Le calcul des références

Comme nous l'avons déjà dit au chapitre 5 (§ 1.2.2.), c'est à partir des valeurs sémantiques (de détermination, de personne, de temps, d'aspect, de modalité,...) que peuvent s'effectuer les calculs référentiels : la langue en effet ne reflète pas passivement le monde, mais elle le représente d'une manière médiatisée. Les calculs référentiels concernent d'une part la réfé-

rence **nominale** (savoir à quel individu du monde renvoie par exemple un syntagme nominal comme *le fils de Jean*), d'autre part la référence **verbale** (savoir à quel événement du monde, situé spatio-temporellement, renvoie par exemple un syntagme verbal comme *est venu hier*).

Les **déictiques** sont les marqueurs privilégiés d'une référence directe au monde : c'est le cas entre autres de pronoms personnels, de possessifs, de démonstratifs, de locutions adverbiales spatiales ou temporelles, etc. La principale difficulté provient du fait que la plupart de ces marqueurs peuvent jouer indifféremment le rôle d'anaphorique ou de déictique (à part quelques notables exceptions : 1re et 2ème personnes du singulier, *hier, aujourd'hui, demain*). C'est encore la conséquence directe des qualités d'évocation du langage : ce qui est construit au fur et à mesure par le texte peut être désigné par les mêmes termes que ce qui existe dans le monde ; ainsi *ce livre-ci* peut signifier aussi bien un livre qui est proche de moi que le dernier livre dont je viens de parler, *C'est lui le coupable* peut se dire aussi bien d'une personne que l'on désigne du doigt, que d'un personnage que l'on vient de "mettre en scène" dans son récit, et même *ici* et *maintenant* peuvent désigner une "position" spatio-temporelle différente de celle de l'énonciation (*Il avait enfin réussi à revenir dans son village natal : maintenant il se sentait réellement en sécurité ; ici plus rien ne pouvait lui arriver...*).

D'une manière générale, au-delà du problème spécifique des déictiques, il n'est pas facile de déterminer quand un élément du discours réfère à un objet nouveau introduit par cet élément, et quand il co-réfère à un objet qui a déjà été évoqué. Comparez par exemple le texte donné en introduction : *Je roulais sur la partie droite de la chaussée quand un véhicule arrivant en face dans le virage a été complètement déporté. Serrant à droite au maximum, je n'ai pu éviter la voiture qui arrivait à grande vitesse,* avec celui-ci : *Je roulais sur la partie droite de la chaussée quand un véhicule arrivant en face dans le virage a été complètement déporté. Serrant à droite au maximum, je n'ai pu éviter la voiture qui me précédait.* Alors que dans le premier texte *qui arrivait à grande vitesse* est un simple qualificatif qui décrit plus précisément le véhicule dont on a déjà parlé, dans le deuxième texte *qui me précédait* est un déterminatif qui spécifie un nouveau véhicule dont il n'a jamais été question précédemment : il est clair que seul un calcul prenant en compte toute la scène peut permettre de différencier ces deux cas de figure.

En fait, mises à part certaines désignations univoques, relativement rares (noms propres, etc.), toute détermination de référence ne peut qu'être le résultat d'un calcul, même dans le cas de formulations qui semblent à première vue non ambiguës, comme par exemple *le directeur du théatre municipal* : en fait la référence dépend de la ville et de la période évoquées, et cela peut ne pas être du tout facile à déterminer, puisque, comme nous l'avons vu, les références spatio-temporelles sont en constante évolution tout au long du texte.

1.3. Le calcul des significations implicites

Un message n'est pas seulement une suite d'informations sur le monde (ou plutôt sur un monde — réel ou imaginaire) qu'enverrait un émetteur en direction d'un récepteur : tout message, même le plus centré sur la fonction informative, référentielle, contient nécessairement des traces du rapport de l'émetteur au contenu de son propre message (même quand on adopte le style neutre du rapport objectif, c'est déjà une façon de se situer par rapport à ce que l'on dit), et aussi des traces de la visée de l'interlocuteur par l'émetteur (même le simple fait de transmettre une information objective à autrui est déjà une façon de tenter de changer quelque chose chez le récepteur, par exemple l'état de ses connaissances).

C'est de cette dimension que tente de rendre compte la théorie des **actes de langage**, élaborée dans le cadre de la pragmatique linguistique. Un acte de langage se définit comme le fait d'effectuer une certaine action, en tant qu'émetteur, en direction du récepteur, par le fait même d'énoncer ce que l'on énonce ; cf. le titre de l'ouvrage de J-L. Austin (1962) : "How to do things with words" ("Quand dire c'est faire"). Le cas des **performatifs** est à cet égard bien connu : quand un président de séance énonce *Je déclare la séance ouverte*, ou quand un prêtre énonce *Je te baptise*, la séance est *ipso facto* ouverte, la personne baptisée ; l'acte est accompli (performé) par la simple profération des paroles correspondantes de la part d'un émetteur investi du pouvoir de le faire (il n'y a pas performatif si ce n'est pas à la 1re personne du singulier et au présent, ou si c'est émis par n'importe qui).

Il existe bien d'autres actes de langage, et l'on parle de **valeurs illocutoires** pour désigner le fait qu'un énoncé donné peut avoir valeur de promesse, de menace, de suggestion, d'ordre, etc. Il y a acte de langage **direct** lorsque cette valeur est explicitement marquée par une forme linguistique (ex : *Je te promets de venir* ; *Je te suggère de partir* ; *Je t'ordonne de répondre*, etc.). Il y a acte de langage **indirect** lorsque la valeur doit être reconstruite, par un mécanisme langagier d'inférence, à partir des formes (ex : *Je viendrai ; Tu ferais mieux de partir ; Pourrais-tu répondre, oui ou non ?*). Le cheminement (la "dérivation") est parfois complexe, et des connaissances pragmatiques fines sont nécessaires pour établir les règles de calcul des actes de langage indirects à partir de la diversité des formes qui peuvent les induire ; ainsi par exemple pour comprendre que l'énoncé *Il fait froid ici* peut être une manière indirecte de demander à quelqu'un qu'il ferme une fenêtre.

La dimension de l'**implicite** est très importante pour la compréhension de la signification complète. La détection des **présupposés** en particulier permet d'épingler ce que l'énonciation fait passer subrepticement comme connu, acquis, évident ou incontestable, sans l'asserter explicitement. Ainsi un énoncé comme *Ton frère continue à fumer* contient deux présupposés, à savoir d'une part que "tu as un (et un seul) frère", et d'autre part qu'"il fumait

déjà auparavant". De façon plus large, on n'appréhende vraiment la signification complète que si l'on est capable de restituer le **non-dit**, d'interpréter pourquoi l'émetteur a dit ce qu'il a dit de la façon dont il l'a dit, bref de faire une série d'inférences sur son intention de signification à partir du dit. Comment savoir, par exemple, qu'un énoncé comme *Tu es belle ce soir*, derrière les apparences de compliment, peut cacher le sous-entendu plus perfide "pour une fois que tu fais un effort de toilette !".

Les systèmes de compréhension automatique de textes, et en particulier de dialogues, rencontrent tous ces problèmes. Pour comprendre, par exemple, qu'un énoncé comme *Je suis en année de maîtrise* peut constituer une réponse appropriée ("pertinente") à la question *Vas-tu au séminaire de X ?*, il faut que le récepteur restitue l'enchaînement "le séminaire de X s'adresse à des étudiants qui sont en année de maîtrise, et donc mon interlocuteur y va, puisqu'il est en maîtrise" (la réponse est alors équivalente à "oui"), ou au contraire l'enchaînement "le séminaire de X s'adresse à des étudiants qui sont dans une autre année que celle de maîtrise, et donc mon interlocuteur n'y va pas" (la réponse est alors équivalente à "non"). De même, pour comprendre la cohérence de l'échange suivant :

X : - *Je n'ai plus de feu, ma chandelle est morte*
Y : - *Va chez la voisine, je crois qu'elle y est* ,

il faut inférer un état de manque à partir du simple constat de X ("je n'ai plus de feu (...)"), comprendre que pour y remédier X essaie d'obtenir que Y lui procure ce qui lui manque, que Y décrypte cette demande indirecte, se refuse à y accéder, et donne seulement à X une information qui lui permettra d'aller chercher ailleurs ce qui lui manque.

La modélisation de phénomènes de ce genre est évidemment complexe. Elle doit faire appel à un certain nombre de notions théoriques que nous nous contenterons ici de mentionner comme par exemple les "maximes conversationnelles" de H. Grice (1975), ou la notion de "pertinence" développée par D. Sperber & D. Wilson (1986). Ces problèmes seront réabordés plus spécifiquement à propos du dialogue dans le chapitre 10 (cf. *infra*) ; pour un développement sur la prise en compte de ces phénomènes en traitement automatique, nous renvoyons le lecteur à G. Sabah (1988 chapitre 10 § 2).

Enfin, par-delà les trois questions qui viennent d'être évoquées dans le présent § 1., il convient de rappeler une dimension importante des textes, qui devrait être prise en compte par les systèmes de compréhension automatique : il s'agit de tout ce qui concerne d'une part l'organisation rhétorique et argumentative du texte, d'autre part les phénomènes de cohérence et de cohésion textuelles ; pour une présentation linguistique de ces problématiques, voir au § 2 des repères bibliographiques donnés à la fin du présent chapitre.

2. Les outils de représentation

Les problèmes que nous venons de passer en revue auront convaincu le lecteur de l'impossibilité d'une "compréhension" automatique sans la mise en place d'outils de représentation permettant d'effectuer les calculs nécessaires à la construction progressive du sens d'un texte. Bien sûr, ces outils peuvent être plus ou moins frustres suivant le type de tâche que l'on veut effectuer ; mais inévitablement se pose le problème de la mise en relation de l'information extraite du texte avec des connaissances extra-linguistiques, dont il nous faut maintenant voir comment elles peuvent être structurées.

2.1. Compréhension et modélisation

Pour pouvoir interpréter correctement des textes, les systèmes de compréhension doivent donc disposer de connaissances de type **encyclopédique**, qui dépassent largement les connaissances "lexicales" dont nous avons parlé au chapitre 5 (§ 2.1.) à propos de l'analyse sémantique de phrases isolées. Même si la frontière entre lexique et encyclopédie comporte une part d'arbitraire, il faut prendre conscience que les connaissances nécessaires ici ne peuvent plus se limiter à ce qui peut raisonnablement entrer dans la définition d'un mot.

Prenons le domaine de la circulation routière et la compréhension des textes de déclaration d'accidents tels que celui que nous avons présenté en introduction. On peut penser qu'un réseau sémantique lexical contiendra des connaissances permettant de savoir qu'une voiture a en général quatre roues, qu'un rétroviseur sert à voir ce qui se passe derrière soi, et à la rigueur qu'un feu rouge est un signal qui interdit de passer. Mais cela n'est pas suffisant : on a besoin aussi de toutes sortes d'informations qui n'ont rien à voir avec le lexique (même si l'on en a une vision très large), comme par exemple de savoir que l'on roule à droite (du moins dans un certain nombre de pays), que quand le feu est vert sur un axe, il est rouge pour l'axe transversal, que quand il pleut, la chaussée devient glissante et qu'il faut s'arrêter au coup de sifflet du gendarme ! Il faut aussi que le système possède, sous une forme ou sous une autre, des connaissances sur l'organisation spatiale des voies de circulation (rues, routes, autoroutes, ponts, tunnels, passages à niveau...) et des notions de cinématique et de dynamique : vitesse, accélération, force centrifuge, dérapage, etc. De plus, pour comprendre certaines formulations dans les déclarations d'accidents, il faut connaître le rôle de ces déclarations et en particulier savoir que le plus souvent l'objectif non avoué du narrateur est d'essayer de se disculper auprès de sa compagnie d'assurance...

Il faut donc envisager une organisation des connaissances qui ne soit plus liée aux mots de la langue, mais plutôt aux objets, aux événements et aux

"lois" du **domaine** que l'on traite. En somme, il faut **modéliser** le domaine en question, et ce que l'on entend par "compréhension d'un texte par la machine" ne peut se définir que par rapport à ce modèle : les performances d'un système de compréhension vont dépendre de manière étroite de ses compétences en matière de modélisation du domaine dans lequel il doit opérer.

Mais si la modélisation du domaine est une condition nécessaire, elle n'est pas suffisante pour assurer la compréhension : en effet, dans les textes que le système aura à analyser, on va forcément être confronté à des formulations dont la richesse dépasse largement le domaine visé. Les exemples de J. Pitrat (1985, pp. 14 *sqq.*) à propos des commentaires de parties d'échecs sont particulièrement frappants de ce point de vue. Ainsi, pour comprendre le sens "échiquéen" de *Les pièces noires sont agglomérées comme un troupeau de moutons sur lequel le loup s'apprête à fondre*, il ne suffit pas d'avoir modélisé l'univers des échecs : il faut un certain nombre de connaissances générales, par exemple sur les animaux (savoir qu'un troupeau de moutons est un ensemble d'animaux mal structuré, que le mouton n'est pas agressif et qu'il ne sait pas se défendre, que le loup peut l'attaquer très rapidement et à coup sûr, etc.).

L'idéal serait bien sûr de pouvoir modéliser l'ensemble des connaissances d'un lecteur humain, mais ceci est nettement hors de portée des systèmes actuels, et pour longtemps : il existe bien des projets grandioses en ce sens, comme par exemple le projet américain CYC de constitution d'un système comportant des centaines de milliers de structures conceptuelles (cf. D. Lenat & R. Guha 1990), mais on peut rester sceptique sur leurs chances de succès. On est donc astreint, quand on conçoit un système de compréhension, à clarifier nettement le statut des connaissances que l'on compte y inclure. D'une part il faut délimiter strictement le domaine que l'on modélise : c'est par rapport à ce modèle que seront construites les représentations des textes analysés, et c'est dans ce cadre que le système doit être opérationnel (répondre à des questions, alimenter une base de données, etc.). D'autre part, il faut ajouter des connaissances linguistiques et extra-linguistiques "générales", dont le but est exclusivement d'aider à l'analyse des textes (compréhension des métaphores les plus courantes, par exemple) ; ces connaissances seront en nombre forcément limité, ce qui implique un compromis et des limites à la capacité du système de comprendre n'importe quel texte dans le domaine visé.

2.2. Les connaissances d'univers

Représenter des connaissances d'univers n'est pas un problème spécifique au traitement automatique des langues : c'est une question centrale de l'Intelligence Artificielle depuis plus de vingt ans (cf. *supra* l'Introduction du présent ouvrage, § 3.3.). Beaucoup de systèmes utilisent, sous des variantes diverses, une forme de représentation qui a été introduite par M. Minsky en 1974 : les "**frames**", que l'on traduit généralement en français par "schémas".

En fait, un schéma est une structure attribut-valeur (cf. *supra* chapitre 5, § 2.1.) dans laquelle on peut associer toutes sortes d'informations à un attribut donné, en particulier des informations procédurales. On peut ainsi spécifier des contrôles à effectuer sur des contraintes que doit vérifier la valeur éventuelle de cet attribut, on peut aussi spécifier des méthodes de calcul de cette valeur si on en a besoin et si elle n'est pas présente, on peut également déclencher d'autres actions sur d'autres éléments du système quand on la modifie, etc. En somme, les systèmes de "frames" sont des réseaux sémantiques auxquels on a donné la capacité d'effectuer localement des calculs spécifiques. On obtient ainsi une grande **souplesse** de représentation de connaissances, puisque l'on peut par exemple disposer simultanément de plusieurs méthodes pour obtenir la valeur d'un attribut : calculer cette valeur à partir d'autres attributs, aller chercher la valeur par défaut d'un prototype, etc. De même, la cohérence du système repose en grande partie sur des **contrôles locaux**, qui permettent de vérifier qu'une valeur est "vraisemblable", de déclencher des actions spéciales dans le cas contraire, etc. Là encore, les techniques informatiques qui permettent d'implémenter ce type de structures sont aujourd'hui très banalisées grâce aux **langages à objets**, dont l'une des caractéristiques essentielles est justement de permettre de décrire des structures (les classes d'objets) définies par des champs (les attributs) et des méthodes spécifiques à ces classes. L'intérêt des "frames" est de pouvoir représenter des classes d'entités et d'événements du monde, auxquels on confère les propriétés que ces entités et événements possèdent habituellement. C'est donc bien un outil de modélisation. L'exemple typique d'utilisation est celui des "scénarios" (ou "scripts") développés par R. Schank (cf. R. Schank & R. Abelson 1977) ; un scénario "*restaurant*" par exemple permet de représenter les événements auxquels on doit s'attendre quand une personne se rend au restaurant : il s'installe, consulte le menu, passe sa commande à un garçon, mange, demande l'addition, paie, etc. Un tel script permet de comprendre un petit texte du type *Nous avions décidé d'aller au restaurant, mais le garçon était d'une lenteur si exaspérante que nous sommes partis avant même d'avoir commandé.*

Des techniques issues de la **logique** ont été aussi beaucoup utilisées, en Intelligence Artificielle en général, pour modéliser des connaissances sur le monde. Au fond, c'est ce principe général de description d'un domaine par un ensemble de règles qui est à l'origine de la vogue des **systèmes experts** ; l'élaboration de systèmes experts est devenu, à l'heure actuelle, une véritable technologie : un nouveau métier est même né de ce type d'activités, celui d'ingénieur "cogniticien", spécialisé dans la mise en forme des connaissances à partir de discussions avec un expert du domaine. En ce qui concerne les systèmes de compréhension, les formalismes logiques les plus utilisés sont les **logiques typées** et les **logiques des défauts**, que nous avons présentées brièvement au chapitre 5, § 2.2. L'avantage essentiel de ces formalismes, c'est qu'ils permettent d'exprimer simplement les "lois" d'un domaine, contrairement aux "frames" qui sont mieux adaptés à la description des objets

du domaine. L'autre grande différence entre ces deux types de formalisme concerne le **contrôle de la cohérence**, qui est d'une importance capitale, dès qu'il s'agit de réaliser une base de connaissances de taille respectable. Alors que ce contrôle est **local** dans les "frames", il est **global** dans les systèmes logiques : c'est ainsi que les logiques des défauts réclament des systèmes dits "de maintenance de la vérité" ("truth maintenance systems") dont le rôle est de détecter les contradictions qui peuvent résulter de l'application des règles de défauts et de rétablir la cohérence en "bloquant" la règle par défaut qui s'est trouvée ... en défaut ! On voit qu'il y a une certaine complémentarité entre les techniques de représentation par "frames" et celles fondées sur la logique. En fait, on assiste aujourd'hui à une convergence entre les deux approches avec l'apparition de systèmes qui cherchent à combiner leurs avantages respectifs. Quoi qu'il en soit, l'obstacle décisif reste celui de la **taille** : il est très facile, dans tous ces formalismes, de construire des exemples "jouets" qui modélisent un petit nombre de connaissances. Mais la réalisation d'un système "en vraie grandeur", même dans un domaine limité, réclame des efforts considérables : le principal problème n'est pas tant le choix du formalisme que la qualité de la conceptualisation du domaine étudié.

Parallèlement à ces outils généraux de représentation, on assiste de plus en plus au développement de modèles "**analogiques**" de phénomènes spécifiques, comme les relations spatiales ou temporelles. On entend par là des modèles dans lesquels la représentation interne reproduit la structure de ce que l'on doit modéliser : un axe pour une structure temporelle, un espace géométrique plan pour des relations spatiales bi-dimensionnelles, etc. Le "raisonnement" sur les connaissances se fait alors en simulant les données sur le modèle. De tels modèles peuvent d'ailleurs être intégrés à un système général : c'est ainsi que dans le projet LILOG (cf *infra*, § 3.), les connaissances spatiales (en l'occurrence la localisation de bâtiments publics dans une ville) sont traitées par deux modules qui coopèrent, l'un utilisant un formalisme logique et l'autre des représentations bidimensionnelles "picturales" ("depictions"). Une des évolutions les plus intéressantes des systèmes de représentation, c'est le développement d'architectures **distribuées**, dans lesquelles les connaissances sont réparties dans différents modules spécialisés qui coopèrent pour fournir une représentation d'ensemble.

2.3. Construction de la représentation interne du texte

D'une manière générale, un texte introduit des entités de diverses natures (personnages, objets, événements,...), les situe dans l'espace et le temps, leur attribue des propriétés de toutes sortes, et décrit les relations qu'elles entretiennent. Comme nous l'avons vu, comprendre un texte revient à en construire une représentation interne, dans le cadre d'une modélisation donnée.

Le premier problème qui se pose est donc celui de repérer l'introduction

d'une **entité nouvelle** et d'en créer une représentation interne, représentation qui évoluera au cours de l'analyse du texte, au fur et à mesure que le texte complètera l'information initiale sur cette entité. Le mécanisme de base qui permet ces créations d'entités nouvelles est l'**instanciation** : on dispose dans le modèle d'une description des classes d'entités susceptibles d'être rencontrées dans le texte, et l'on crée un élément nouveau appartenant à l'une de ces classes (une instance de la classe) qui représentera l'entité en question tout au long du texte. Si l'on reprend l'exemple de la déclaration d'accident automobile donné en introduction, on sera ainsi amené à créer des représentations associées aux deux véhicules, aux deux conducteurs, à la route, au virage, et aux événements correspondant aux faits que les deux véhicules se déplacent, que l'un d'entre eux est déporté, qu'il y a collision, etc. Bien entendu, cela suppose qu'il existe dans la base de connaissances, sous une forme ou sous une autre, les classes "*véhicule*", "*conducteur*", "*route*", "*collision*", etc. L'omniprésence des anaphores rend cette opération très délicate. En fait le système doit constamment résoudre la question : s'agit-il d'une nouvelle entité ou bien d'une entité déjà créée ? Les règles qui régissent la réponse à ces questions jouent un rôle décisif et, répétons-le, elles ne sont pas simples.

Le deuxième problème est d'établir les **relations** qui existent entre ces diverses entités : au premier chef, d'une part, les relations entre les différents acteurs des événements et d'autre part les relations spatio-temporelles entre événements. Dans l'exemple de l'accident, il s'agit d'établir les relations entre les conducteurs, leur voiture, les événements "*déplacement*", et la route ; d'établir aussi que le virage est une portion de la route, qu'un des événements "*déplacement*" a lieu dans le virage, que les deux mouvements ont lieu en sens opposé (et l'un vers l'autre) ; de déterminer qui est déporté, où a lieu la collision ; et bien sûr d'établir la succession temporelle de ces événements. On le voit, certaines de ces relations sont explicites dans le texte, au sens où les données linguistiques suffisent à les obtenir, tandis que d'autres doivent être déduites par un raisonnement qui utilise des connaissances extra-linguistiques contenues dans la base. Ainsi la relation entre le premier conducteur, sa voiture et le premier événement "*déplacement*" est contenue dans la forme linguistique *je roulais*, alors que l'existence du deuxième conducteur et la relation de celui-ci avec le deuxième véhicule ne peuvent être déduites qu'en faisant intervenir une règle (par défaut) du domaine du type "un véhicule qui roule sur une route a en général un conducteur", puisque rien dans le texte même n'y fait référence. De même, la formulation *arrivant en face* permet d'établir la relation spatiale entre les déplacements des deux véhicules, alors que pour localiser la collision, il faut faire appel à un raisonnement s'appuyant sur des connaissances solides en cinématique.

La question qui se pose est donc de savoir quand et comment déclencher les **raisonnements** extra-linguistiques nécessaires à la représentation complète du texte dans le cadre du modèle. Deux attitudes s'opposent à ce sujet.

La première consiste à ne pas opérer de distinction entre connaissances linguistiques et extra-linguistiques, et donc à utiliser toutes les règles applicables au fur et à mesure que se construit la représentation du texte. La difficulté est alors de savoir comment **contrôler** le processus de déduction, pour rester dans des limites raisonnables, puisqu'*a priori* un nombre faramineux de faits peuvent être ainsi obtenus (s'il est question d'une voiture qui se déplace, on peut en déduire bien sûr l'existence d'un conducteur, mais aussi d'un volant, de roues, d'un rétroviseur, etc.). Une solution élégante est proposée à ce problème par la théorie de la "profondeur variable" développée par D. Kayser (1987) : à chaque fait déduit est associé un "niveau de profondeur" et le système cherche à travailler à profondeur constante, autant que possible. En particulier, la création d'une nouvelle entité augmentant le niveau de profondeur, le système n'y procèdera (lors de la résoultion d'une anaphore, par exemple) que si l'attribution des nouvelles propriétés à l'une des entités existantes conduit à une contradiction : c'est le principe dit "circonscriptif".

La deuxième attitude consiste à séparer nettement les deux activités d'analyse du texte et de modélisation proprement dite : on produit à partir du texte une représentation des entités, propriétés et relations qui sont directement explicitables à partir des formes linguistiques, et c'est un autre module qui prend ensuite en entrée ces données et utilise les connaissances extra-linguistiques ; les raisonnements qu'effectue ce module sont alors généralement guidés par la tâche qu'il a à accomplir (réponse à une question, etc.). On évite ainsi la prolifération des déductions dans la mesure où le système de raisonnement cherche à atteindre des buts précis, et non pas à donner une représentation "complète" du texte.

Les deux formalismes "généralistes" les plus utilisés pour analyser les textes ont été introduits dans le chapitre 5 (cf. *supra*). Il s'agit de la théorie de la représentation du discours (DRT) de H. Kamp et de la théorie des graphes conceptuels de J. Sowa. En effet, l'ambition de ces deux formalismes, que nous avons présentés à propos de l'analyse sémantique de la phrase, dépassent largement ce cadre : l'une de leurs principales qualités est justement de permettre de traiter des phénomènes inter-phrastiques comme l'anaphore, en se donnant les moyens de représenter des cadres de référence qui permettent le calcul des co-références et, dans une certaine mesure, des relations spatio-temporelles. Ceci étant, il faut reconnaître que ces moyens ne sont encore que très frustres par rapport à la complexité des phénomènes dont il faut rendre compte, en particulier dans le domaine de l'anaphore associative, des références temporelles, ...

3. Les architectures de systèmes de compréhension

Voyons maintenant où en est la réalisation de systèmes de compréhension.

3.1. Des maquettes aux systèmes opérationnels

Pendant longtemps, les systèmes de compréhension automatique n'ont été que des maquettes de laboratoire, fort utiles pour le développement de la recherche, mais ne donnant pas lieu à des applications en vraie grandeur. En effet, et ceci est commun à la plupart des secteurs de l'intelligence artificielle, il est assez facile de réaliser des maquettes de petite taille qui peuvent impressionner par des résultats spectaculaires, mais il est beaucoup plus difficile de passer à la taille supérieure : la maîtrise du comportement du système devient bien plus délicate, en particulier dès qu'il n'est plus possible d'assurer "manuellement" le contrôle de la cohérence des bases de connaissance (cf. *supra*, § 2.2.). Si l'on ajoute à cela une tendance bien compréhensible des industriels impliqués dans ce domaine à surestimer les performances de leurs futurs produits, même quand ils ne sont encore qu'à l'état de maquettes, on comprend qu'une évaluation objective des progrès réels ne soit vraiment pas simple, avec d'un côté des travaux de recherche fort passionnants, mais limités dans leurs ambitions, et de l'autre des annonces à grand renfort de publicité de l'arrivée sur le marché de produits mirifiques qui tiennent rarement leurs promesses.

On peut malgré tout penser que cette période d'immaturité est aujourd'hui pratiquement dépassée : en effet, on assiste à la réalisation de systèmes qui sont beaucoup plus modestes dans leurs objectifs, tout en mettant en oeuvre des moyens d'une ampleur considérable. Très souvent, ces systèmes proviennent d'une collaboration entre chercheurs universitaires et industriels, et ils sont toujours le fruit de projets à long terme, réclamant pour réussir (comme dans tous les gros projets informatiques) des investissements conséquents et une planification d'une très grande rigueur.

Plutôt que de tenter de donner une vue d'ensemble qui serait forcément parcellaire et incomplète, nous allons présenter deux exemples typiques de réalisation en insistant sur les facteurs qui ont contribué à en faire des succès.

3.2. Un système de petites annonces

Le système PALME, système de petites annonces d'offres d'emploi du journal *Le Monde*, réalisé par GSI-ERLI (cf. J. Vega 1990) est très typique d'un système opérationnel d'alimentation et d'interrogation d'une base de données classique. Il s'agit d'un service télématique vidéotex (accessible par Minitel : 3615 LM), permettant au grand public d'accéder aux offres parues dans le journal, soit en posant une question en formulation libre, soit en décri-

vant son curriculum vitae. Le système sélectionne alors les offres d'emplois adaptées au profil du demandeur. Le système comporte deux grandes parties :

- Une indexation automatique, à partir des annonces parues dans le journal. Chaque offre d'emploi est analysée par le système et archivée dans la base de données avec une indexation qui s'appuie sur un réseau sémantique : ainsi une annonce telle que *Cherche ingénieur pour coordonner une équipe chargée de vendre nos SGBD* sera indexée par les trois descripteurs *responsable*, *vente*, et *base de données*. Il faut tout de même noter que cette indexation automatique reste sous le contrôle des équipes éditoriales du journal, qui la vérifient et peuvent le cas échéant la rectifier.

- Le système d'interrogation proprement dit, qui consiste à analyser le CV rempli (ou la question posée) par l'utilisateur du service Minitel, et à classer les offres d'emploi disponibles par proximité sémantique avec la demande. L'analyse du CV s'effectue suivant les mêmes principes que l'analyse d'une offre d'emploi, et l'adéquation d'un CV et d'une offre est évaluée à partir d'une "distance" entre les descripteurs du CV et ceux de l'offre d'emploi. Cette distance est calculée sur le réseau sémantique, en fonction de la proximité dans le graphe des descripteurs.

Un tel système est fabriqué à partir de la "boîte à outils" de GSI-ERLI (dénommée ALETH), qui est constituée de modules linguistiques et extra-linguistiques : un dictionnaire de 70 000 mots (il s'agit en fait de la partie française de GENELEX, programme européen de constitution de bases de données lexicales, cf *supra*, ch 3 § 4), comportant des dizaines de milliers de descriptions syntaxiques et de relations sémantiques ; une gamme d'outils syntaxiques, systèmes de règles bien sûr, mais aussi des modules très spécifiques, comme l'identification de dates, le redressage orthographique (existence d'un phonétiseur), etc. ; enfin un outil de confection et de révision de réseaux sémantiques, qui est le coeur de la partie "intelligente" du système.

Ces outils permettent à GSI-ERLI de fabriquer des applications de ce type relativement facilement (une réalisation comme PALME représente tout de même un investissement de quelques 4 MF). Il faut noter que cette entreprise constitue une expérience assez originale dans ce secteur (tout au moins en Europe) : avec ses 15 ans d'existence, elle a pu accumuler progressivement une expérience faite à la fois de travaux de recherche (en liaison avec les milieux universitaires) et de développements applicatifs lui permettant de se confronter de manière réaliste aux besoins. En 1993, avec plus de 50 personnes dans le département "ingénierie linguistique et documentaire", c'est sans doute le premier spécialiste européen des industries de la langue.

3.3. Un système d'informations touristiques

Parmi les gros projets qui ont été entrepris au cours de ces dernières années dans le domaine de la compréhension automatique de textes, le projet

LILOG ("Linguistic and Logic methods and tools") du département "recherche et développement" d'IBM-Allemagne est en tous points exemplaire. Lancé en 1985, son objectif était de construire une chaîne complète d'outils suffisamment généraux pour être utilisables dans toutes sortes d'applications dans ce domaine. Dès le départ, la décision (sage !) a été prise de tester les performances d'un tel atelier de développement sur une application particulière : c'est un système d'informations touristiques sur une grande ville (Düsseldorf) qui a été choisie. Le système devait pouvoir :

- lire les brochures touristiques de la ville et intégrer l'information pertinente dans sa base de connaissances, qui devait donc contenir des informations historiques (date de construction d'un monument, par exemple), géographiques (localisation des édifices) et pratiques (heures d'ouverture et prix d'entrée d'un musée, etc.) ;

- dialoguer en allemand avec un utilisateur (analyse et génération donc) dans le but de le renseigner sur les activités touristiques dans la ville : dire non seulement quand le musée est ouvert et ce que l'on peut y trouver, mais aussi comment y aller (itinéraires en bus, à pied, etc.), sachant où se trouve l'utilisateur.

Comme on le voit ce projet est beaucoup plus ambitieux qu'un simple système d'interrogation de base documentaire : il y a en particulier des problèmes de représentation spatiale et temporelle de grande ampleur, qui contraignent le système à une compréhension en profondeur des textes et des questions qui lui sont soumis.

Le projet a duré 5 ans 1/2, il a impliqué en tout près de 200 personnes (en moyenne 60 personnes par an, soit plus de 4000 hommes-mois). En plus des laboratoires d'IBM, cinq universités allemandes y ont participé. Une organisation très modulaire avec une planification rigoureuse a été mise en place. Dans un domaine où les vastes projets impliquant plusieurs groupes débouchent rarement sur des réalisations concrètes, LILOG constitue un succès assez remarquable. En effet, toute l'architecture a été mise en place (plus de 200 000 lignes de code !) et l'application-test tourne dans des conditions satisfaisantes, bien qu'encore expérimentales (en particulier, les textes de brochures touristiques qui servent à alimenter la base de connaissance doivent subir quelques transformations "manuelles" avant d'être traitées par le système).

Le système comporte trois grandes parties : un analyseur, un système de manipulation de connaissances, et un générateur. L'analyseur se décompose à son tour en trois grandes composantes de niveau morphologique-lexical (avec un analyseur morphologique particulèrement élaboré pour les mots composés, le traitement des mots inconnus, etc.), un analyseur syntaxico-sémantique (basé sur HPSG pour la syntaxe, cf *supra* chapitre 4 §3.3, et réalisant une première étape de l'analyse sémantique dite "compositionnelle"), un module de traitements sémantiques complémentaires (les représentations sémantiques sont inspirées de la DRT, cf. *supra* chapitre 5 § 2.2). Le système de manipulation de connaissances est fondé sur un formalisme de logique typée, qui

intègre beaucoup d'idées issues des réseaux sémantiques (héritage et défauts, en particulier). Un gros effort a été fourni aussi pour une gestion efficace de grandes bases de connaissances. On trouvera dans G. Herzog & C.R. Rollinger (1991) une présentation assez détaillée de toutes les composantes de LILOG.

4. Perspectives

Les deux systèmes que nous venons de présenter sont assez représentatifs des capacités actuelles en matière de compréhension automatique de textes : comme on l'a vu avec le premier exemple, on est donc capable aujourd'hui de traiter des aspects limités du sens d'un texte tout-venant de manière à pouvoir par exemple l'indexer correctement dans une base de données. Par ailleurs, on peut (mais cela réclame des moyens considérables) aller plus loin pour obtenir une compréhension plus profonde, là encore limitée à un domaine très précis. Il est tout à fait significatif de noter la prudence et la modestie avec laquelle les responsables de LILOG s'expriment quand ils font le bilan de leur projet : ils sont tentés de reprendre la formule de Neil Armstrong à propos de sa marche sur la Lune ("un petit pas pour un homme, un grand bond pour l'humanité"), mais en l'inversant (*op. cit.* p.29) : l'énorme effort investi dans ce projet ne constitue qu'un tout petit pas pour la recherche, car l'essentiel reste encore à faire.

Cette attitude tranche beaucoup avec le ton triomphaliste qui dominait à la fin des années 70, où, dans le sillage des travaux de R. Schank en particulier, apparaissaient les premières maquettes de systèmes capables de traiter des textes tout-venant pour alimenter des bases de données (ou pour les traduire) : ainsi, les présentations de systèmes capables, à partir de dépêches d'agence de presse, d'alimenter des bases de données sur les catastrophes naturelles, les attentats terroristes ou encore les déplacements d'hommes d'État, pouvaient donner l'impression que les réalisations opérationnelles n'allaient pas tarder. Or, plus de dix ans après, il faut bien constater qu'aucun système professionnel d'envergure basé sur ces principes ne fonctionne aujourd'hui dans le secteur de la presse.

C'est ce changement d'attitude qui nous semble le phénomène le plus porteur d'espoir pour les progrès dans le domaine dans les prochaines années : en effet, la lucidité devant la complexité de la tâche est une condition décisive pour mener à bien des travaux qui, du coup, ne peuvent s'envisager que dans des perspectives pluri-disciplinaires (avec en particulier une forte composante linguistique), et des collaborations étroites entre la recherche fondamentale et la recherche appliquée, et aussi (ce n'est pas forcément la même chose) entre les laboratoires universitaires et les entreprises spécialisées.

Catherine FUCHS et Bernard VICTORRI
(ELSAP-CNRS)

Repères bibliographiques

1. Problématiques générales de la compréhension automatique :

ANDREEWSKY, E. (ed.) (1985) : "Langage et cognition", *T.A. Informations*, Paris, Kincksieck, 1985 : 1.

[Recueil d'articles sur la compréhension et le traitement automatique.]

ALLEN, J. (1987) : *Natural language understanding*, Menlo Park, Benjamin / Cummings.

[Ouvrage synthétique sur la compréhension.]

KAYSER, D. (1985) : Des machines qui comprennent notre langue, *La Recherche*, 170 : 16, Paris, 1198-1212.

[Article de vulgarisation sur la compréhension automatique.]

PITRAT, J. (1985) : *Textes, ordinateurs et compréhension*, Paris, Eyrolles.

[Ouvrage de base très accessible et très vivant.]

SABAH, G. (1989) : *L'intelligence artificielle et le langage* (vol. 2 : "Processus de compréhension"), Paris, Hermès.

[Voir en particulier le ch. 7 : "Inférences", le ch. 8 : "Références" et le ch. 9 : "Modélisation de textes".]

SABAH, G. (1992) : Lecture : peut mieux faire, *Sciences et Avenir*, n° 86 hors-série, Paris, 32-37.

[Article de vulgarisation.]

SMITH, G. (1991) : *Computers and human language*, Oxford, Oxford University Press.

[Voir le ch. 11 : "Discourse interpretation using world knowledge" et le ch. 12 : "Knowledge about discourse".]

2. Pragmatique et linguistique :

ARMENGAUD, F. (1985) : *La pragmatique*, Paris, P.U.F., Coll. Que sais-je ?

[Introduction aux courants logico-philosophiques fondateurs de la pragmatique.]

AUSTIN, J-L. (1962) : *How to do things with words*, Oxford, Oxford University Press. Trad. fr. (1970) : *Quand dire, c'est faire*, Paris, Seuil.

[L'un des ouvrages "pionniers" en matière d'actes de langage.]

BLACK, J. & WILENSKY, R. (1979) : An evaluation of story grammars, *Cognitive Science*, 3, Norwood N.J., Ablex, 213-230.

[Critique des approches de type "grammaires de récits".]

DUCROT, O. (1988) : L'argumentation dans la langue : bibliographie, *Modèles Linguistiques*, , X : 2, Lille, 131-132.

FUCHS, C. & LE GOFFIC, P. (1992) : *Les linguistiques contemporaines ; repères théoriques*, Paris, Hachette.

[Voir en particulier le ch. 11 : "Enonciation et pragmatique" : présentation du schéma de la communication et des fonctions du langage selon Jakobson (1963) ; du courant énonciatif (Benveniste) sur les indiciels, les modalités et la structuration des énoncés ; du courant pragmatique sur les présupposés et l'implication du sens, la théorie des actes de langage et l'interaction communicative.]

GRICE, H. (1975) : Logic and conversation, dans Coles & Morgan (eds.) : *Syntax and semantics 3 : speech acts*, New-York, Academic Press, 41-58.

[Article de référence sur les "maximes conversationnelles".]

REICHENBACH, H. (1947) : *Elements of symbolic logic*, New-York, Macmillan.

[Contient en particulier les bases d'une logique du temps.]

SEARLE, J. (1969) : *Speech acts ; an essay in the philosophy of language*, Cambridge, Cambridge University Press.

[L'un des classiques "historiques" de la théorie des actes de langage.]

SPERBER, D. & WILSON, D. (1986) : *Relevance, communication and cognition*, Oxford, Blackwell ; trad.fr. (1989) *La pertinence : communication et cognition*, Paris, Minuit.

[L'ouvrage de référence sur la question de la "pertinence".]

STRAWSON, F. (1970) : Phrase et acte de parole, *Langages*, 42, Paris, Larousse, 19-33.

[Réflexion de pragmatique linguistique sur les différents niveaux de la signification.]

VAN DIJK, T. (1985) : *Handbook of discourse analysis*, Londres, Academic Press, 4 volumes.

VAN DIJK, T. & KINTSCH, W. (1983) : *Strategies of discourse comprehension*, New-York, Academic Press.

[Les deux ouvrages ci-dessus sont consacrés à l'étude des mécanismes de discours.]

Revue *Modèles Linguistiques*, Lille, n° IX : 1 (1987) : "Aspects de la cohésion et de la cohérence discursives", et n°X : 2 (1988) : "Analyse transphrastique".

3. Formalismes et implémentations :

BRADY, M. & BERWICK, R. (eds.) (1983) : *Computional models of discourse*, Cambridge Mass., M.I.T. Press.

[Recueil de contributions sur la modélisation du discours en vue du traitement automatique.]

HERZOG, G. & ROLLINGER, C.R. (ed.) (1991) : *Text understanding in LILOG*, Berlin, Springer, Coll. "Lecture notes on artificial intelligence".

[Le projet LILOG, système de développement d'applications de compréhension automatique. Voir en particulier pp. 703-718 l'application-test sur le domaine des informations touristiques et pp. 719-733 un bilan de la conduite du projet]

JAYEZ, J. (1988) : *L'inférence en langue naturelle*, Paris, Hermès.

[Etude des mécanismes inférentiels en langue "naturelle".]

KAMP, H. (1974) : A theory of truth and semantic representation, dans J. Groenendijk *& al.* (eds.) : *Truth, interpretation and information*, Dordrecht, Foris.

["Théorie de la représentation du discours" (DRT).]

KAYSER, D (1987) : "Représentation du sens ou représentation des connaissances", *Actes du Colloque "Sémantique formelle ; fondements philosophiques et applications"*, Grenoble — version anglaise : "Meaning representation versus knowledge representation", dans N. Cooper & P. Engel (eds) (1991) : *New inquiries into meaning and truth*, New York, St Martins Press, 163-186.

[Théorie de la profondeur variable.]

LENAT, D. & GUHA R. (1990) : *Building large knowledge based systems : representation and inference in the CYC project,* Reading Mass. Addison Wesley.

[Projet américain CYC d'un système de connaissances encyclopédiques.]

MINSKY, M. (1974) : *A framework for representing knowledge*, Cambridge Mass., M.I.T.

[Représentation par "frames".]

SABAH, G. (1990) : CARAMEL : un système multi-experts pour le traitement automatique des langues, *Modèles Linguistiques*, XII : 1, Lille, 95-118.

[Prototype de laboratoire développé au LIMSI d'Orsay.]

SCHANK, R. & ABELSON, R. (1977) : *Scripts, plans, goals and understanding*, Hillsdale, Erlbaum.

[Représentation par "scripts" ("scenarios").]

SOWA, J. (1984) : *Conceptual structures : information processing in mind and machine*, New-York, Addison Wesley.

[Théorie des "graphes conceptuels".]

VEGA J. (1990) : Semantic matching between job offers and job search requests, *Actes du colloque COLING 90*, Helsinki, 67-69

[Présentation de l'application PALME, système automatisé d'offres d'emploi.]

NB : Voir aussi les repères bibliographiques de notre chapitre 5 ("Sémantique").

9
GÉNÉRATION AUTOMATIQUE DE TEXTES

Dans la communication entre l'homme et la machine en langue naturelle, le domaine de la génération s'occupe de produire les réponses de la machine dans la langue souhaitée par l'usager. Après un bref survol de l'évolution du domaine, ce chapitre sera divisé en deux parties qui suivent la modularisation d'un système de génération. La première partie traitera de la détermination du contenu informatif du texte produit par la machine. La seconde traitera de la formulation de ces informations dans une langue naturelle (étant entendu que le texte ainsi construit doit se situer dans un registre de langue considéré comme "soutenu"). Notre présentation sera principalement centrée sur les difficultés linguistiques recontrées lors de la réalisation d'un système de génération. Pour un exposé synthétique des recherches menées en génération, nous renvoyons à l'article très complet de M. Zock & G. Sabah (1992), qui fourmille de références bibliographiques.

1. Repères historiques

Le domaine de la génération de textes est beaucoup plus récent que celui de l'analyse et de la compréhension. De nombreux programmes d'analyse ont vu le jour depuis les années 1950, tandis que les premiers programmes de génération n'ont surgi que vers la fin des années 1970. Cette dissymétrie entre analyse et génération peut s'expliquer par le fait que la machine a toujours produit des textes, par le biais de **textes pré-enregistrés**, comme les messages systèmes du type *Assurez-vous que l'imprimante est bien connectée*, dont l'affichage à l'écran est géré par le système d'exploitation de la machine.

Cette méthode, qui évite la réalisation d'un système de génération, n'a pas d'équivalent en analyse. Un texte adéquat étant pré-enregistré pour chaque cas de figure offert par l'application, la production d'une réponse de la machine consiste simplement à rechercher dans l'ensemble des textes pré-enregistrés celui correspondant à la situation concernée. Cette méthode est tout à fait satisfaisante lorsque le nombre de cas de figure pour l'application envisagée est restreint. Sinon, des problèmes de mémoire et de recherche surgissent : la place occupée par un grand nombre de textes pré-enregistrés peut être rédhibitoire, tout comme la recherche d'un texte pré-enregistré dans un grand ensemble de tels textes.

La méthode des textes pré-enregistrés peut être améliorée en utilisant des **variables**. Par exemple, on peut pré-enregistrer le message suivant : *Il n'est pas possible d'effacer votre catalogue ? x qui n'est pas vide*, où la variable *? x* correspond à un nom de catalogue. La méthode des textes pré-enregistrés à variable est efficace et fréquemment utilisée lorsque les variables correspondent à des valeurs numériques ou à des noms propres. Mais elle trouve vite ses limites dès que l'on cherche à élargir la portée des variables (ce qui est le cas dans les applications ambitieuses visées dans les années 90) : les variables ne correspondant pas à des valeurs numériques ou à des noms propres obligent à multiplier le nombre de textes pré-enregistrés et à les annoter d'informations syntaxiques (cf. L. Danlos 1990) ; on se trouve en fait confronté aux difficultés de la génération, alors que le but initial était de les éviter.

Les premiers systèmes de génération n'avaient pas pour objectif la production de textes dans un acte de communication donné. Il s'agissait simplement d'engendrer des phrases à contenu informatif aléatoire, le but étant de tester des théories syntaxiques. Arrivèrent ensuite des systèmes qui ne pouvaient engendrer que des **phrases isolées**. Ce n'est donc qu'au début des années 80 qu'on aborda le problème de la génération de **textes** dans un acte de communication donné. Dans ce cadre, le texte généré par la machine doit satisfaire deux exigences : d'une part indiquer à l'utilisateur les informations qu'il désire, et d'autre part offrir une formulation de ces informations dans une langue correcte. Il s'ensuit que le processus de génération comporte **deux composants** : le premier (système expert de raisonnement) traite la question "Quoi dire ?" (détermination du contenu informatif), le second (module de génération linguistique) traite la question "Comment le dire ?" (formulation du contenu informatif dans une langue correcte).

Ces deux composants seront présentés dans les paragraphes suivants. Mais auparavant quelques remarques :

- Dans les applications du traitement automatique du langage naturel qui demandent à la fois un module d'analyse et un module de génération (par exemple interrogation de base de données, traduction bidirectionnelle L <—> L'), un courant de recherches s'intéresse aux modules **réversibles** : l'analyse et la

génération partagent les mêmes bases de données linguistiques, les algorithmes sont symétriques (cf. *supra*, chapitre 7 § 4.1).

- Des recherches en génération sont menées sur la réalisation de systèmes **psychologiquement plausibles** simulant la production langagière d'un locuteur humain. Ces recherches débouchent sur des modèles "incrémentaux", c'est-à-dire des modéles où les questions "Quoi dire ?" et "Comment le dire ?" ne sont pas traitées séquentiellement comme présentées ci-dessous mais parallèlement. Elles ne seront pas abordées ici, où nous nous intéressons à la production de textes écrits rédigés dans un style soutenu et traduisant fidèlement les informations de la représentation sémantique fournie en entrée.

- La génération automatique de textes écrits peut être prolongée par la synthèse de messages **oraux**. Que ce soit dans le cas d'un système d'interrogation de bases de données, de production de commentaires sur des données numériques, ou de traduction automatique, les réponses de la machine peuvent être orales et / ou écrites. La production de messages oraux demande d'interfacer un système de génération avec un système de synthèse de la parole. Cette interface doit calculer la représentation phonétique des mots et la prosodie des phrases (cf. *supra*, chapitres 1 et 2).

2. Le contenu informatif (la question "Quoi dire ?")

Dans le présent paragraphe, nous allons examiner, à travers les principales applications de la génération, en quels termes se pose la question "Quoi dire ?". Le § 3 sera consacré à la question "Comment le dire ?".

Considérons tout d'abord le cas d'un système d'**interrogation d'une base de données** dans le cadre de la communication homme-machine en langue naturelle (qui sera examinée plus en détail au chapitre 10) : ce cas correspond à ce qui est représenté dans la Figure I.

Ex : Entrée : *Comment être à Cluny le 27 Août dans l'après- midi en partant de Paris le matin ?*

Sortie : *Vous pouvez prendre un train à 10h15 pour Macon, puis un bus pour Cluny.*

Le module de raisonnement accède à la base de données pour y trouver les informations demandées par l'usager. Lorsque l'interrogation de la base de données fait partie d'un dialogue (suite de couples question-réponse), le module de raisonnement gère aussi un historique du dialogue afin de tenir compte des réponses déjà transmises et des réponses qui ne satisfont pas complètement l'usager.

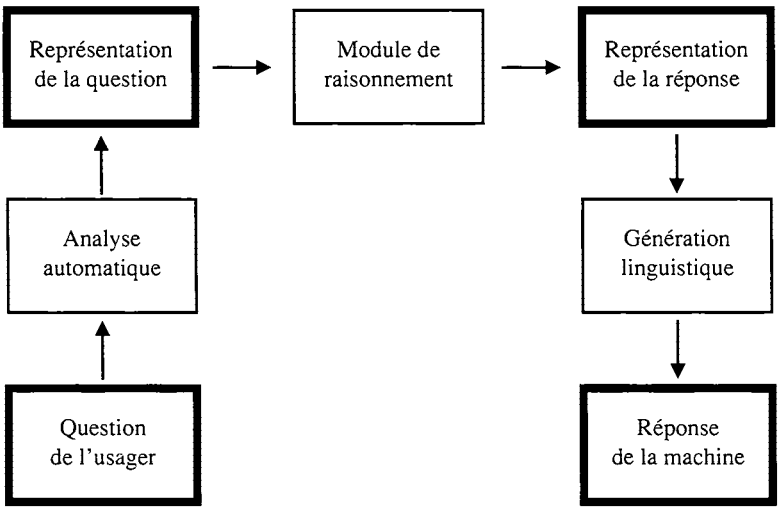


FIGURE I
Système d'interrogation d'une base de données

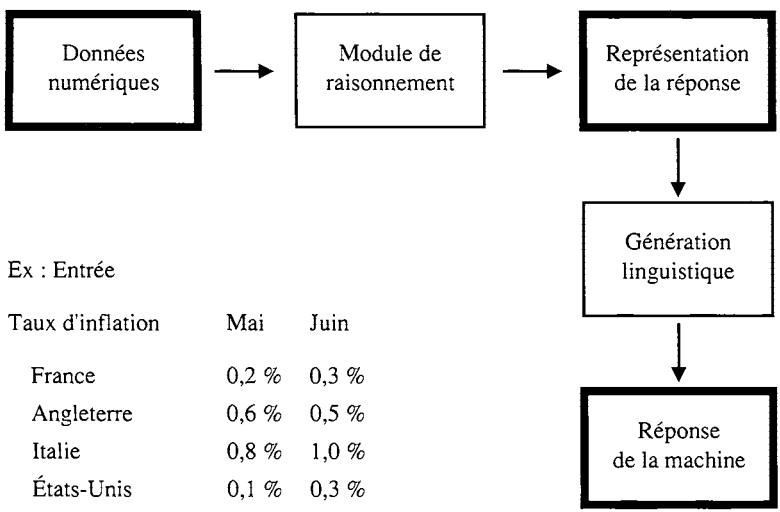


FIGURE II
Système commentant des données numériques

Les systèmes produisant des **commentaires sur des données numériques** ne requièrent pas, quant à eux, de module d'analyse automatique, car les données numériques sont directement interprétables par la machine (cf. Figure II).

Sortie : *Le taux d'inflation a augmenté ce mois-ci : il est à 0,3% alors qu'il était à 0,2% le mois dernier. Cette tendance à la hausse s'est aussi observée aux Etats-Unis et en Italie. Par contre, l'Angleterre a enregistré une baisse de son taux d'inflation.*

Que ce soit dans le cadre d'une interrogation de base de données ou dans celui de production de commentaires sur des données numériques, le module de raisonnement doit fournir une représentation de la réponse qui tende à un échange **coopératif** (selon les termes de H. P. Grice : voir *infra*, chapitre 10) ; en d'autres termes, la réponse doit satisfaire au maximum l'usager.

Les recherches menées sur la question "Quoi dire ?" pour fournir la représentation d'une réponse coopérative se situent principalement dans le domaine des **sciences cognitives**. Par contre, les recherches menées sur la question "Comment le dire ?" se situent dans le domaine de la **linguistique computationnelle**.

Au carrrefour de ces deux domaines, se situent les recherches sur la représentation des connaissances et la caractérisation d'une **représentation sémantique**. La question de la représentation sémantique se pose aussi dans un système de traduction (cf. Figure III et *supra*, chapitre 7).

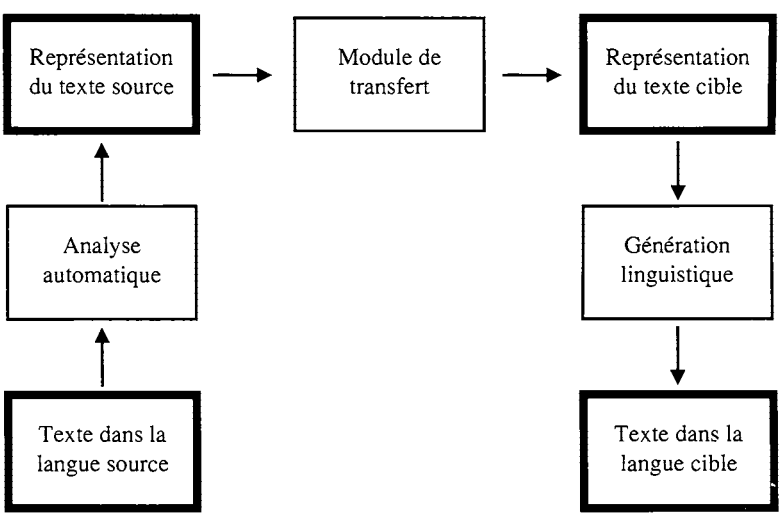


FIGURE III
Système de traduction automatique

Ex : Entrée : *Mary was told that John left yesterday*

Sortie : *On a dit à Marie que Jean était parti hier.*

Comme on l'a vu au chapitre 7 (voir *supra*), dans un système de traduction, le module de raisonnement est remplacé par un module de transfert : la détermination du contenu informatif du texte dans la langue-objet ne se pose pas, puisque ce contenu informatif doit, en principe, être identique à celui du texte dans la langue-source. Par contre, un module de tranfert s'impose pour passer de la langue-source à la langue-objet. La quantité de travail effectuée par le module de génération dépend du degré d'abstraction des représentations intermédiaires : plus ces représentations sont abstraites (plus ces représentations sont sémantiques et s'éloignent des formes de surface), plus se posent les questions inhérentes à la génération de texte, qui seront présentées dans les paragraphes suivants.

La question de la caractérisation d'une **représentation sémantique** est donc essentielle, tant en traduction qu'en génération. Néanmoins, il faut bien dire qu'à l'heure actuelle il n'existe guère de consensus sur une telle représentation. Des outils et des formalismes ont été développés pour manipuler des représentations sémantiques, mais le contenu de ces représentations varie encore trop d'un auteur à l'autre. Pour la génération, ceci signifie que les entrées des générateurs ne constituent pas des entités stables.

Toutefois, pour avancer une représentation sémantique qui servira de point de départ aux exemples de génération linguistique des paragraphes suivants, nous proposerons les principes suivants : une représentation sémantique contient un certain nombre de **concepts** (notés C_i) liés entre eux par des **relations** (notées $\neg$). Ces relations, qui sont binaires, ternaires ou n-aires, sont d'ordre rhétorique (par exemple relation d'exemplification $C_i \neg C_j$ où C_i est un exemple de C_j) ou d'ordre sémantique (par exemple relation causale ou temporelle). On peut considérer une représentation sémantique comme un concept $C_0 = C_1 \neg C_2 \neg \ldots \neg C_n$. Les concepts C_i sont soit des prédicats (notés $PRED_i$), soit récursivement un ensemble de concepts liés entre eux par des relations : C_i = PREDi / $C^1_i \neg C^2_i \neg \ldots \neg C^m_i$. Les **prédicats** sont des foncteurs liant entre eux des entités ou des prédicats. Les **entités** représentent les objets concrets et les êtres animés. Les entités et les prédicats sont organisés selon une hiérarchie de classes génériques propres au domaine d'application du générateur. Le **domaine d'application** d'un générateur se définit comme le type de textes qu'il produit, par exemple résumé de match de football, rapport boursier, bulletin météorologique.

La Figure IV illustre la représentation d'un prédicat écrite dans un formalisme "orienté objet", formalisme qui sera adopté dans la suite de ce chapitre.

Le prédicat PRED1 est une instance de la classe **ORDRE** qui est caractérisée par trois champs : **AGENT**, dont la valeur est l'étiquette *LOCUTEUR*, **INTERLOCUTEUR** dont la valeur est l'étiquette MARIE, et **OBJET** dont la

PRED1 =: **ORDRE**
AGENT : *LOCUTEUR*
INTERLOCUTEUR : MARIE
OBJET : PRED2

PRED2 =: **TRANSACTION**
ACHETEUR : JEAN
VENDEUR : MARIE
MARCHANDISE : OBJET1

OBJET1 =: **OBJET**
TYPE : maison
DÉFINITUDE : défini
NOMBRE : 1

LOCUTEUR =: **HUMAIN**
TITRE : *Locuteur*
SEXE : masculin

MARIE =: **HUMAIN**
NOM : Marie
SEXE : féminin

JEAN =: **HUMAIN**
NOM : Jean
SEXE : masculin

FIGURE IV
Représentation d'un prédicat

valeur est l'étiquette PRED2 de la classe **TRANSACTION**. Cette classe est elle-même caractérisée par trois champs : **ACHETEUR** (valeur JEAN), **VENDEUR** (valeur MARIE) et **MARCHANDISE** (valeur OBJET1).

3. La génération linguistique (la question "Comment le dire ?")

Il est généralement admis que la génération linguistique se traite par une série de **modules**, le premier prenant les décisions conceptuelles (par exemple ordre des informations), les suivants prenant les décisions linguistiques (par exemple choix lexicaux et choix des constructions syntaxiques), l'avant-dernier appliquant les règles de la syntaxe, le dernier les règles de la morphologie. Cette modularisation repose sur les hypothèses suivantes :

1) les opérations de "haut niveau" doivent être effectuées avant les opérations de "bas niveau" ;

2) les opérations conceptuelles sont de haut niveau, les opérations de choix lexical sont d'un niveau moyen, les opérations syntaxiques sont de bas niveau, et les opérations morphologiques sont de très bas niveau.

En analyse, on a admis des hypothèses exactement symétriques jusque dans les années 80 ; elles ont été remises en question et les systèmes actuels ne reposent plus sur une approche stratificationnelle (cf. *supra*, chapitres 3 à 5).

On arrive en génération au même type de conclusion : une approche stratificationnelle ne permet pas de prendre en compte les nombreuses dépendances entre les différents niveaux. Dans ce paragraphe, nous commencerons par présenter l'approche classique, c'est-à-dire l'approche stratificationnelle

et récursive, ce qui nous permettra d'exposer les principales décisions qui doivent être prises dans un système de génération. Ensuite, nous mettrons en évidence différentes dépendances entre les niveaux, qui nous amèneront à avancer une approche non modulaire.

3.1. Approche récursive et stratificationnelle

Considérons l'organisation d'ensemble et les différentes opérations, hiérarchiquement ordonnées dans l'approche classique.

3.1.1 Présentation

Reprenons la formalisation d'une représentation sémantique C_0 présentée au § 2 :

$$C_0 = C_1 \neg C_2 \neg \ldots \neg C_n$$

avec C_i = PREDi (concept "simple")/

$C^1_i \neg C^2_i \neg \ldots \neg C^m_i$. (concept "complexe")

La synthèse d'un concept complexe C_i demande d'effectuer deux opérations : d'une part, déterminer dans quel ordre doivent apparaître les formes synthétisant les concepts C_i et comment relier entre elles ces formes avec la sémantique des relations $\neg$ (opération connue sous le nom de "sélection d'une structure de discours"), d'autre part, déterminer comment synthétiser C_i. Dans une approche récursive et stratificationnelle, ces deux opérations sont effectuées dans un **ordre fixe** : en premier lieu, on sélectionne une structure de discours, ensuite on appelle récursivement l'algorithme de synthèse d'un concept.

Prenons comme exemple le concept $C_O = C_1 \neg C_2 \neg C_3$ et supposons que la structure de discours sélectionnée consiste à exprimer C_2 en premier, puis C_1, et enfin C_3 introduit par le connecteur *mais*. La synthèse de C_0 notée SYN (C_0) donne donc :

$$SYN\ (C_0) = SYN\ (C_2)\ .\ SYN\ (C_1).\ Mais\ SYN\ (C_3).$$

Supposons maintenant que les concepts C_i soient tous des concepts simples, c'est-à-dire des prédicats $PRED_i$ et examinons la synthèse de ces prédicats. Un prédicat peut être synthétisé en une phrase ou en un groupe nominal complexe. Par exemple, le prédicat PRED2 présenté dans la section précédente peut être exprimé, entre autres, de l'une des façons suivantes :

(1) *Marie a vendu la maison à Jean*

(2) *Jean a acheté la maison à Marie*

La vente de la maison à Jean par Marie

L'achat de la maison à Marie par Jean.

Néanmoins, par souci de simplification, nous nous contenterons d'examiner la synthèse d'un prédicat en une phrase. Une phrase repose sur un "élé-

ment prédicatif", en (1) le verbe *vendre*, en (2) le verbe *acheter* (l'élément prédicatif d'une phrase n'est pas nécessairement le verbe : ainsi dans une construction à copule, par exemple). La synthèse d'un prédicat demande donc une opération de choix lexical, c'est-à-dire choix de l'élément prédicatif de la phrase. Il faut aussi déterminer la construction syntaxique de l'élément prédicatif : en (1) et (2) les verbes sont construits à l'actif, mais on pourrait choisir de les construire au passif, par exemple. Enfin, il faut synthétiser les compléments de l'élément prédicatif et appliquer les règles de la morpho-syntaxe. Dans une approche stratificationnelle, ces opérations sont effectuées séquentiellement. Reprenons notre exemple : $C_0 = PRED_1 \neg PRED_2 \neg PRED_3$. Sa synthèse s'effectue en franchissant les étapes suivantes :

- 1° étape : sélection d'une structure de discours, ce qui fournit :

  SYN (C_0) = SYN ($PRED_2$) . SYN ($PRED_1$) . Mais SYN ($PRED_3$).

Puis successivement, pour chaque prédicat PRED\S(j;i) :

- 2ème étape : choix lexical de l'élément prédicatif
- 3ème étape : choix d'une construction syntaxique
- 4ème étape : synthèse des compléments
- 5ème étape : application des règles de la morpho-syntaxe.

Cette approche est séduisante car elle permet un **algorithme modulaire** où interviennent séquentiellement sémantique, syntaxe et morphologie. Elle se heurte néanmoins à des obstacles linguistiques qui seront exposés ci-dessous. Avant d'aborder ces obstacles, examinons comment doivent s'effectuer la sélection d'une structure de discours et les choix lexicaux pour les éléments prédicatifs.

3.1.2. Sélection d'une structure de discours

Un texte n'est pas une suite de phrases agencées au hasard. Un texte possède une structure qui est, en premier lieu, déterminée par le type d'informations qu'il véhicule. Que ce soit un résumé de match de football, un rapport boursier ou un bulletin météorologique, les informations s'organisent toujours selon un nombre limité de **schémas stéréotypés**. L'ensemble des schémas stéréotypés d'un type de texte donné constitue une "grammaire de discours". Cette base de données linguistiques doit être intégrée dans le générateur élaboré pour le domaine d'application en question.

On peut être tenté de rapprocher la notion de "schémas stéréotypés" utilisée en génération de la notion de "scripts" ou "scénarios" utilisée en analyse, ou plutôt en compréhension de texte (voir *supra* chapitre 8). Néanmoins, il n'y a guère de rapport entre ces deux notions : un schéma stéréotypé décrit une structure textuelle tandis qu'un scénario décrit un enchaînement d'événements de la vie réelle.

Une des opérations qu'un générateur doit effectuer pour produire un texte du domaine d'application consiste donc à choisir un élément dans cette base de données linguistiques, opération connue sous le terme de "sélection d'une structure de discours". La **sélection d'une structure de discours** est déterminée principalement par les informations contenues dans l'entrée du générateur. A titre d'illustration, considérons un générateur produisant des bulletins météorologiques (cf. R. Kittredge *& al.* 1986). Les bulletins météorologiques peuvent s'organiser selon les schémas suivants (très simplifés) :

- Schéma 1 : 1) description de la carte du ciel et des éventuelles précipitations,
  2) description de la température,
  3) description du vent,
  etc.
- Schéma 2 : 1) annonce d'une tempête de vent,
  2) description des éventuelles précipitations,
  etc.
- Schéma 3 : 1) annonce d'une vague de chaleur,
  2) description de la carte du ciel,
  etc.

Lorsque l'entrée du générateur (des données numériques triées et regroupées par le système expert en charge de la question "Quoi dire ?") correspond à un jour "normal", le schéma 1 est choisi. Les schémas 2 et 3 sont respectivement choisis en cas de tempête de vent ou en cas de vague de chaleur.

La sélection d'une structure de discours n'est pas une opération qui doit être effectuée récursivement. Pour montrer ce point, considérons un concept complexe qui représente une relation causale entre deux concepts notés CAUSE et RESULTAT. Une structure de discours possible pour exprimer un tel concept est :

(SD1) SYN (CAUSE), et de ce fait SYN (RESULTAT).
= *Luc a renversé son verre, et de ce fait il a sali la nappe.*
= *Luc est malade, et de ce fait Marie est de mauvaise humeur.*

Mais cette structure de discours donne des résultats désastreux si, par exemple, le concept CAUSE est lui-même un concept complexe qui représente une conjonction de deux événements, le discours suivant est maladroit :

Luc est malade et Jean est parti, et de ce fait Marie est de mauvaise humeur.

Pour un tel cas, il vaut mieux choisir, par exemple, une structure de discours utilisant la conjonction *parce que* :

(SD2) SYN (RESULTAT) parce que SYN (CAUSE).

= *Marie est de mauvaise humeur parce que Luc est malade et que Jean est parti.*

Mais cette structure de discours donnerait à son tour des résultats désastreux si le concept CAUSE était lui-même une relation causale exprimée au moyen de (SD2) :

Marie est de mauvaise humeur parce que Luc est malade parce qu'il a trop mangé.

Ces exemples montrent que la sélection d'une structure de discours ne doit pas se faire récursivement mais **globalement** : la sélection d'une structure de discours pour un concept complexe $C_1 \neg C_2 \neg \ldots \neg C_n$ demande de prendre en compte les structures de discours disponibles pour chaque concept complexe C_i .

Pratiquement tous les chercheurs en génération s'accordent pour intégrer une **grammaire de discours** dans les données linguistiques d'un générateur. Par contre, le type d'informations représenté varie d'un chercheur à l'autre. Les chercheurs qui écrivent un algorithme récursif et stratificationnel n'incluent que des informations conceptuelles, par exemple type d'informations et ordre de présentation de ces informations, comme dans la version simplifiée des schémas de bulletins météorologiques présentés ci-dessus (cf. W. Man & J. Moore 1981 ; K. McKeown 1985). Cependant, nous montrerons (voir *infra*, § 3.2.1) qu'il est préferable d'y intégrer des informations linguistiques.

3.1.3. Choix lexicaux pour les éléments prédicatifs

Le générateur inclut une base de données lexicales qui associe à chaque prédicat manipulé dans le domaine d'application la ou les façons de l'exprimer. A titre d'illustration, un prédicat de la classe **ORDRE** peut être exprimé par le "schéma de phrase" suivant (les catégories :dir-obj et :à-obj sont respectivement des abréviations d'objet direct et d'objet indirect introduit par *à*) :

P1 (:P (:sujet ?AGENT) (:verbe *ordonner*) (:dir-obj ?OBJET)
(:à-obj ?INTERLOCUTEUR))

où ?AGENT, ?OBJET et ?INTERLOCUTEUR désignent des variables qui doivent être instanciées respectivement par les valeurs des champs **AGENT, OBJET** et **INTERLOCUTEUR** du prédicat de la classe **ORDRE**. De même, un prédicat de la classe **TRANSACTION** peut être exprimé par un des schémas de phrase suivants :

P2a (:P (:sujet ?ACHETEUR) (:verbe *acheter*) (:dir-obj ?MARCHANDISE)
(:à-obj ?VENDEUR))
b (:P (:sujet ?VENDEUR) (:verbe *vendre*) (:dir-obj ?MARCHANDISE)
(:à-obj ?ACHETEUR))

Lorsque plusieurs possibilités sont offertes pour un même prédicat, une des opérations que le générateur doit effectuer consiste à choisir la meilleure, opération connue sous le terme de "choix lexical". Le choix d'un élément

pour un prédicat donné est déterminé par des considérations "locales" (la définition du prédicat) et non-locales (le contexte du prédicat). Reprenons la représentation (RS1) de la Figure IV.

PRED1 =: **ORDRE**
AGENT : *LOCUTEUR*
INTERLOCUTEUR : MARIE
OBJET : PRED2

PRED2 =: **TRANSACTION**
ACHETEUR : JEAN
VENDEUR : MARIE
MARCHANDISE : OBJET1

OBJET1 =: **OBJET**
TYPE : maison
DÉFINITUDE : défini
NOMBRE : 1

LOCUTEUR =: **HUMAIN**
TITRE : *Locuteur*
SEXE : masculin

MARIE =: **HUMAIN**
NOM : Marie
SEXE : féminin

JEAN =: **HUMAIN**
NOM : Jean
SEXE : masculin

FIGURE IV
Représentation d'un prédicat

et considérons le prédicat PRED2 instance de **TRANSACTION**. Tel qu'il est défini, le choix entre P2a et P2b est localement indifférent : ces deux schémas de phrase sont aussi valables l'un que l'autre. Ceci ne serait plus le cas, par exemple, pour une instance de **TRANSACTION** dont le champ **ACHETEUR** ne serait pas spécifié et dont le champ **MARCHANDISE** serait indéfini : le schéma P2b serait localement préférable à P2a dans la mesure où il permet de générer une phrase comme *Marie a vendu du lait* qui est plus naturelle qu'une phrase générée à partir de P2a telle que *Du lait a été acheté à Marie*. Examinons les considérations non-locales : PRED2 est enchâssé dans PRED1 et ce fait doit être pris en compte. En effet, selon le choix de P2a ou P2b, le générateur va produire une des formes suivantes :

(a) *J'ai ordonné à Marie que Jean lui achète la maison.*

(b) *J'ai ordonné à Marie de vendre la maison à Jean.*

La forme (b) construite avec une infinitive dont le sujet est coréférent à *Marie* est préférable à la forme (a) construite avec une complétive. Pour engendrer (b), et donc choisir (P2b) et non (P2a), il faut que la sélection d'un schéma de phrase exprimant PRED2 enchâssé dans PRED1 soit assujettie à la condition suivante : choisir (si possible) un schéma de phrase dont la valeur du sujet soit égale à la valeur du à-objet de la principale.

La condition que nous venons d'énoncer repose sur des connaissances **lexicales** : 1) le verbe *ordonner* se construit avec une complétive qui se réduit à une infinitive lorsque son sujet est égal au contrôleur de *ordonner*, c'est-à-dire le à-obj, 2) le verbe *ordonner* se construit plus naturellement avec une complétive qu'avec une infinitive (pour pratiquement tous les verbes se construisant et avec une complétive et avec une infinitive, la forme

avec infinitive est plus naturelle ; il existe cependant des exceptions, comme *annoncer* : *Jean a annoncé (qu'il avait fait une bêtise + ? avoir fait une bêtise*). Ces informations sur la syntaxe d'un verbe doivent être explicitement indiquées dans le générateur, et il en est de même de toute autre information lexicale. En d'autres termes, un générateur doit inclure un "**lexique-grammaire**" qui indique les propriétés syntaxiques de chaque item lexical utilisé dans le domaine d'application. Parmi les propriétés syntaxiques pertinentes figurent, entre autres, les possibilités de "transformations" comme le passif : pour des raisons stylistiques, sémantiques ou syntaxiques, il peut être préférable d'utiliser une construction passive, si le lexique-grammaire indique que cela est possible.

3.2. Interdépendance entre les décisions

L'approche stratificationnelle ne tient pas compte des différentes dépendances entre les niveaux sémantique, syntaxique et lexical. Pourtant, ces dépendances sont nombreuses, comme nous allons l'illustrer ci-dessous par quelques exemples.

3.2.1. Sélection d'une structure de discours et choix des constructions syntaxiques

Considérons les deux discours suivants :

(1) *Mes chemises ont été froissées parce qu'elles ont été entassées dans le tiroir par Luc.*

(2) *Mes chemises ont été froissées par Luc parce qu'elles ont été entassées dans le tiroir.*

Leurs formes ne diffèrent que par la position du complément d'agent *par Luc* qui est dans la subordonnée en (1) et dans la principale en (2). Mais leur sémantique est radicalement différente : le discours (1) signifie que l'entassement des chemises par Luc est la cause directe de leur état, tandis que le discours (2), dans la mesure où il a un sens, signifie que l'entassement des chemises par quelqu'un d'autre que Luc est un motif (ou cause indirecte) qui a poussé Luc à froisser les chemises. Si l'on veut générer un discours qui a la sémantique de (1), c'est-à-dire la sémantique d'une relation causale directe, il faut donc respecter la condition suivante : la conjonction *parce que* ne doit pas introduire une subordonnée au passif sans agent lorsque la principale mentionne l'agent (un discours où l'agent n'est mentionné ni dans la principale ni dans la subordonnée comme *Mes chemises ont été froissées parce que elles ont été entassées dans le tiroir*, induit une sémantique de relation causale directe). Si les strucures de discours ne contiennent que des informations conceptuelles, ce type de condition demande de gérer des dépendances entre la sélection d'une structure de discours et le choix des constructions syntaxiques, ce qui n'est pas évident. Une façon élégante de résoudre le pro-

blème consiste à assortir les structures de discours d'informations linguistiques en indiquant **explicitement** les constructions syntaxiques, comme dans :

SYN (RESULTAT [passif-sans-agent]) parce que SYN (CAUSE [passif-avec-agent]) .

(1) = *Mes chemises ont été froissées parce qu'elles ont été entassées dans le tiroir par Luc.*

Le fait que la structure de discours

SYN (RESULTAT [passif-avec-agent]) parce que SYN (CAUSE [passif-sans-agent]) .

(2) = *Mes chemises ont été froissées par Luc parce qu'elles ont été entassées dans le tiroir.*

ne fasse pas partie de la grammaire de discours d'une relation causale directe empêche de générer un texte comme (2) lorsque l'on veut obtenir une sémantique de relation causale directe.

3.2.2. Sélection d'une structure de discours et choix lexicaux

Nous avons vu que les choix lexicaux demandaient de prendre en compte des considérations non locales telles que l'enchâssement d'un prédicat dans un autre. Il existe de plus une dépendance entre les choix lexicaux et la sélection d'une structure de discours, comme nous allons le montrer.

Pour exprimer une relation causale directe, les deux structures de discours suivantes sont possibles :

(SD3) SYN (CAUSE [actif]) . SYN (RESULTAT [actif]) .

(SD4) SYN (RESULTAT [actif]) . SYN (CAUSE [actif]) .

comme en témoignent respectivement les deux discours suivants :

(3a) *Des anarchistes ont fait sauter la voiture du pape. Ils l'ont tué.*

(4a) *Des anarchistes ont tué le pape. Ils ont fait sauter sa voiture.*

Mais si l'on utilise le verbe *assassiner* au lieu du verbe *tuer*, seul un discours de structure (SD4) est autorisé :

(3b) **Des anarchistes ont fait sauter la voiture du pape. Ils l'ont assassiné.*

(4b) *Des anarchistes ont assassiné le pape. Ils ont fait sauter sa voiture.*

La sélection d'une structure de discours et le choix d'items lexicaux sont donc des opérations dépendantes l'une de l'autre : si l'on choisit d'abord la structure (SD3), on ne peut plus choisir *assassiner*, et si l'on choisit d'abord le verbe *assassiner*, on ne peut plus choisir (SD3). Néanmoins, il n'est pas possible d'attribuer une priorité à l'une de ces opérations plutôt qu'à l'autre : dans certains cas, on préfèrera privilégier la sélection d'une structure de discours, dans d'autres cas les choix lexicaux.

3.2.3. Choix lexicaux et constructions syntaxiques

Une construction passive, par exemple, ne peut être utilisée que si le verbe le permet (information fournie par le lexique-grammaire) :

Cette valise est lourde parce qu'elle contient du plomb

**Cette valise est lourde parce qué du plomb est contenu par elle.*

Il existe donc une dépendance entre les choix lexicaux et les choix de constructions syntaxiques.

3.3. Pour une approche non modulaire

Reprenons les trois premières étapes qui sont franchies dans une approche récursive et stratificationnelle :

- 1re étape : sélection d'une structure de discours,
- 2ème étape : choix lexical de l'élément prédicatif pour chaque prédicat,
- 3ème étape : choix d'une construction syntaxique pour chaque élément prédicatif.

Nous venons de voir que toutes ces étapes sont **dépendantes** les unes des autres sans qu'aucune hiérarchie s'impose. Il est donc impossible de les traiter séquentiellement.

On est alors amené à concevoir un algorithme de génération le moins modulaire possible. Le minimum de modularisation est obtenu avec **deux** modules dans le modèle présenté à la Figure V.

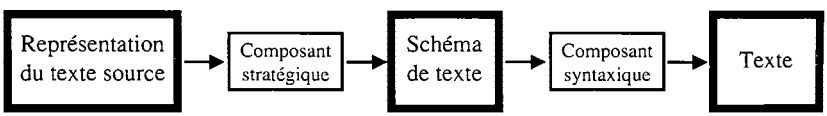


FIGURE V
Modèle de génération à deux modules

Dans ce modèle, le composant **stratégique** sélectionne les structures de discours, les éléments prédicatifs exprimant les prédicats et les constructions syntaxiques des éléments prédicatifs, en respectant l'interdépendance entre ces opérations à l'aide d'heuristiques basées sur le domaine d'application. Il s'appuie sur deux bases de données linguistiques : une grammaire de discours qui inclut des informations conceptuelles et linguistiques, et une base lexicale qui est accompagnée d'un lexique-grammaire. Le composant **syntaxique** applique les règles de morpho-syntaxe à un schéma de texte. Il s'appuie sur une grammaire qui formalise les règles de la morpho-syntaxe et leur application.

La grammaire est la seule base de données linguistiques qui est indépendante du domaine d'application du générateur : les autres bases de données

linguistiques, c'est-à-dire grammaire de discours et base lexicale assortie d'un lexique-grammaire, sont dépendantes du domaine et doivent être élaborées sur la base d'un corpus de textes du domaine. Notons que le lexique-grammaire des items lexicaux concernés par l'application peut lui aussi être considéré comme dépendant du domaine, dans la mesure où une construction syntaxique pour un item donné peut être naturelle ou non selon le type de texte où elle apparaît.

Ce modèle suppose que les opérations effectuées dans le composant syntaxique n'ont pas d'incidence sur celles du composant stratégique. Ceci est vrai pour l'application des règles de la morpho-syntaxe (par exemple la règle d'accord entre un adjectif et un nom doit être respectée quels que soient les éléments en jeu). Toutefois ceci n'est plus vrai pour les opérations de pronominalisation (cf. F. Namer 1990 ; L. Danlos 1992). Celles-ci relèvent par certains aspects de la morpho-syntaxe (par exemple calcul de la forme d'un pronom en fonction de sa position syntaxique et de son genre et nombre). C'est pourquoi elles sont effectuées dans le composant syntaxique. Mais elles mettent aussi en jeu des considérations interphrastiques qui relèvent du composant stratégique. Elles devraient donc influencer les opérations effectuées pour construire un schéma de texte, ce sujet étant l'objet de recherches actuelles. En d'autres termes, les questions de pronominalisation mettent en cause la séquentialité composant stratégique - composant syntaxique. A plus forte raison, elles mettent en cause l'extrême modularité de l'approche récursive et stratificationnelle.

4. Perspectives

Les sytèmes de génération permettent de déboucher sur des **produits fiables,** c'est-à-dire avec un pourcentage d'erreurs nul : en effet, l'entrée d'un système de génération est maîtrisable dans la mesure où c'est la machine qui "a la main". Cette situation contraste avec celle des systèmes d'analyse automatique qui ne peuvent guère déboucher sur des pourcentages d'erreurs nuls dans la mesure où c'est l'usager qui a la main : on ne sait (heureusement) pas maîtriser les idées d'un humain ou sa façon de s'exprimer.

Les enjeux sont donc importants pour les **industries de la langue**. Les premières applications industrielles concernent la génération de bulletins météorologiques (domaine bien délimité qui constitue une des applications industrielles de la traduction automatique) et la production automatique de courrier. On peut s'attendre à ce que les années à venir voient naître de nombreuses autres applications.

Au niveau de la **recherche** en traitement automatique du langage naturel, la génération offre des perspectives nouvelles et prometteuses, surtout pour

les phénomènes de discours que l'on ne peut ignorer si l'on veut produire des textes cohérents, bien rédigés et accomplissant les buts de la communication.

Les difficultés dans la recherche sur le discours résident dans le fait qu'il ne s'agit pas de poser de simples jugements de grammaticalité mais qu'il faut systématiquement aborder le sens du discours, en prenant en compte que la moindre variation peut entraîner un changement de sens radical. Pour illustrer cette dernère affirmation, considérons les deux discours suivants :

(1) *Le Professeur a puni l'élève. Il lui avait lancé des boulettes.*

(2) *Le Professeur a puni l'élève. Il lui a lancé des boulettes.*

Ils ne diffèrent que par le temps du verbe de la seconde phrase : plus-que-parfait en (1), passé composé en (2). Mais leur sens est radicalement différent : en (1) c'est l'élève qui a lancé des boulettes au professeur, tandis qu'en (2) c'est le professeur qui a lancé des boulettes à l'élève.

On se heurte donc à un problème de combinatoire : les paramètres intervenant dans l'interprétation d'un discours sont si nombreux qu'il est difficile de les contrôler tous simultanément. Il est donc nécessaire de mettre au point des méthodes où la prudence est de règle, c'est-à-dire où on ne fait varier les paramètres que l'un après l'autre et doucement ! Heureusement, l'ordinateur vient à l'aide : c'est le générateur en état de production qui fournit des exemples de discours avec une sémantique inadéquate, évitant au chercheur la recherche de tels exemples.

Laurence Danlos
(Université de Paris 7, TALANA)

Repères bibliographiques

1. Monographies sur la génération automatique :

[Les monographies suivantes présentent les travaux des auteurs sur la génération automatique. Elles introduisent à tous les aspects de cette problématique, et ne demandent pas de connaissance préalable sur le domaine :]

DANLOS, L. (1985) : *Génération automatique de textes en langues naturelles*, Paris, Masson.

DANLOS, L. (1987) : *The linguistic basis of text generation*, Cambridge, Cambridge University Press.

MCKEOWN, K. (1985) : *Text generation*, Cambridge, Cambridge University Press.

NOGIER, J-F. (1991) : *Génération automatique de langage et graphes conceptuels*, Paris, Hermès.

[Outre ces monographies, le lecteur non-spécialiste pourra également consulter la référence suivante, où sont réunis divers articles sur la génération de textes, les uns du point de vue du traitement automatique, les autres dans une perspective plus cognitive :]

ANIS, J. (ed.) (1992) : "La génération de textes", *Langages*, 106, Paris, Larousse.

2. Recueils d'articles sur la génération :

[Les livres suivants, classés chronologiquement, ont été publiés à la suite de congrès internationaux ou européens sur la génération. Ils sont organisés en chapitres qui couvrent les différents aspects de la génération. Chaque chapitre comporte plusieurs articles qui expriment les différents points de vue des chercheurs. Ces livres demandent d'être familiarisés avec le domaine :]

KEMPEN, G. (ed) (1987) : *Natural language generation : new results in artificial intelligence, psychology and linguistics*, Dordrecht, Martinus Nijhoff Publishers.

ZOCK, M. & SABAH, G. (eds.) (1988) : *Advances in natural language generation : an interdisciplinary perspective*, Londres, Pinter, & Norwood, Ablex.

MELLISH, C. & ZOCK, M. (eds.) (1990) : *Current research in natural languge generation,* New York, Academic Press.

PARIS C., *& al.* (eds.) (1990) : *Natural language generation in artificial intelligence and computational linguistics*, Dordrecht, Kluwer Academic Press.

3. Etat de l'art en génération :

[Les articles suivants présentent l'état de l'art en génération ou un survol des recherches menées dans le domaine. L'article de M. Zock et G. Sabah fourmille de références bibliographiques :]

DANLOS, L. (1990) : Génération automatique de textes en langues naturelles : état de l'art, *Actes du colloque "Industries de la langue"*, Montréal.

MANN, W. & MOORE, J. (1981) : Computer Generation of Multiparagraph English Text, *American Journal of Computational Linguistics*, Cambridge, M.I.T. Press. 7 : 1.

ZOCK, M. & SABAH, G. (1992) : La génération automatique de textes : trente ans déjà ou presque, *Langages*, 106, Paris, Larousse, 8-35.

4. Génération automatique de bulletins météorologiques :

[L'article suivant illustre la réalisation d'une application de la génération :]

KITTREDGE, R. *& al.* (1986) : Synthesizing Weather Forecast from Formatted Data, *Proceeding of COLING*, Bonn, ACL, 563-565.

5. Pronominalisation en génération :

[Les ouvrages suivants portent sur un problème délicat en génération : la pronominalisation. Ils sont consacrés aux langues romanes dont les systèmes de pronoms pré-verbaux demandent une approche "globale" et non "séquentielle" :]

DANLOS, L. (1992) : Contraintes syntaxiques de pronominalisation en génération de textes, *Langages*, 106, Paris, Larousse, 36-62.

NAMER, F. (1990) : *Pronominalisation et effacement du sujet en génération automatique de textes en langues romanes*, Thèse de Doctorat, Université de Paris 7.

6. Dialogue :

[Pour l'ouvrage suivant, cf. *infra*, la bibliographie du chapitre 10 :]

BILANGE, E. (1992) : *Le dialogue naturel avec une machine*, Paris, Hermès.

7. Génération de messages oraux :

[Pour une présentation d'un système de synthèse de la parole interfacée avec un système de génération, montrant les différences avec la problématique de la synthèse de la parole à partir d'un texte écrit, nous renvoyons aux trois références suivantes, déjà mentionnées dans les chapitres 1 et 2 (cf. *supra*) :]

DANLOS, L. *& al.* (1986) : Synthesis of spoken messages from semantic representations (Semantic-representation-to-speech-system), *Proceedings of COLING 1986*, Bonn, ACL, 599-604.

DANLOS, L. *& al.* (1985) : Synthèse de messages oraux à partir d'une représentation sémantique, *Actes du GALF, 14èmes Journées d'Etudes sur la Parole*, Paris, GALF, 118-121.

EMERARD, F. (1977) : *Synthèse par diphones et traitement de la prosodie*, Thèse de 3ème cycle, Université de Grenoble.

10
DIALOGUE HOMME-MACHINE

Nous préférons parler de **dialogue homme-machine** plutôt que de **communication homme-machine,** car cette dernière recouvre un domaine beaucoup plus vaste, qui intègre notamment la communication non verbale, et qui n'est pas loin de s'identifier avec l'utilisation-même de l'ordinateur. Notre propos se limite en effet à l'interaction avec la machine par le biais du langage, ce qui suppose, outre l'utilisation de l'ordinateur à l'aide de la voix ou de l'écrit, que celui-ci, d'un simple outil, devienne un véritable interlocuteur.

Un système de dialogue homme-machine comporte, outre un module d'analyse (cf. *supra*, chapitre 8) et un module de génération (cf. *supra*, chapitre 9), un module de **raisonnement** spécifique (voir Figure I).

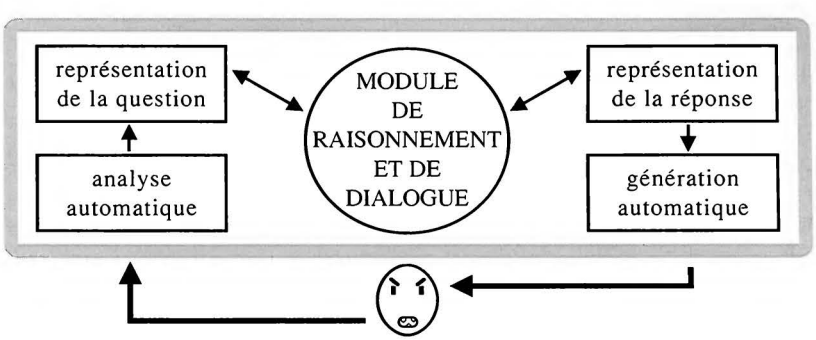


FIGURE I

Raisonner en l'occurrence suppose certes la capacité d'accéder à une base de connaissances quelconque (horaire des trains ou des avions, pages jaunes de l'annuaire, catalogue de vente par correspondance...), de façon à être capable de répondre à des questions, mais cela suppose surtout la capacité de **gérer un dialogue,** c'est-à-dire une séquence indéterminée d'échanges, ce qui revient à être capable de mémoriser, de changer de sujet, de s'adapter à son interlocuteur... Module de raisonnement et module de dialogue sont en sorte très souvent confondus.

Dans un premier temps, nous allons exposer les problèmes spécifiques du dialogue homme-machine. Nous examinerons ensuite les problèmes que posent les modélisations et les mises en oeuvre informatiques.

1. Les problèmes spécifiques du dialogue homme-machine

Le dialogue homme-machine est souvent considéré comme une sous-partie de la **compréhension automatique** du langage, et les problèmes sont fréquemment les mêmes que ceux que l'on rencontre pour l'analyse automatique de textes, notamment en ce qui concerne la prise en considération du contexte. Toutefois, comme nous essaierons de le montrer, le dialogue pose des problèmes spécifiques, au travers de son rapport à la norme notamment : en situation interactive, l'efficacité de la communication l'emporte sur toute autre considération. Par ailleurs, le concept de sens doit être manipulé de façon particulière. Quoi qu'il en soit, nous verrons que le domaine du dialogue homme-machine reste asez restreint.

1.1. Dialogue et compréhension automatique du langage

Au sein des traitements automatiques du langage, dialogue homme-machine et compréhension automatique du langage rencontrent partiellement les même problèmes, car l'un et l'autre sont confrontés au problème du **sens**. Comme on l'a vu au chapitre 6, les applications actuelles des traitements automatiques de la parole ont rarement à prendre en compte le sens. Pour des applications telles que la dictée vocale, la déroute des approches fondées sur les connaissances au profit d'approches d'inspiration stochastique (en l'occurrence l'utilisation de méthodes à base de probabilités de succession) ne fait qu'illustrer le caractère peu linguistique des options retenues. Pour le dialogue en revanche, qu'il soit oral ou écrit, il est absolument inenvisageable de concevoir le développement d'un système sans une maîtrise du sens.

1.1.1. La signification intentionnelle

Le problème central, à la fois en ce qui concerne l'analyse et le dialogue, est ici le problème des **buts**. Il s'agit là non pas d'un terme usuel, mais d'un concept informatique qui a donné lieu à des modélisations spécifiques. Ce terme désigne une réalité simple, à savoir que pour comprendre un énoncé on ne peut se satisfaire du **sens propositionnel**, et qu'il est nécessaire d'accéder en première instance au contexte, et notamment à la **signification intentionnelle** (cf. *supra*, chapitre 8).

Soit l'échange suivant :

H1 : je cherche un garage
M1 : *où cela ?*
H2 : y a-t-il un hôtel à Martinville ?

Comprendre cet échange suppose non seulement que le système sache décoder qu'il s'agit d'une question, que Martinville est un nom de lieu, qu'un hôtel loue des chambres, que dans des chambres on peut dormir, etc., mais également qu'il soit à même d'**inférer** à partir de H2 quel est le **but** poursuivi par l'interlocuteur, en l'occurrence qu'on lui indique un garage, si possible proche d'un hôtel, dans la mesure où il suppose que la nécessaire immobilisation de son véhicule sera trop longue pour lui permettre de repartir dans la journée.

D'aucuns objecteront sans doute que cette remarque sur les buts est une remarque linguistique banale, dans la mesure où cela revient à dire que le signe linguistique tire sa signification tout autant de son utilisation que de son référent. Mais l'intérêt des traitements automatiques du langage est justement de ne pas distinguer le banal de l'original et de mettre surtout en évidence l'important. En dialogue homme-machine, cette question des buts est capitale, à telle enseigne que la formulation, c'est-à-dire le sens propositionnel, est secondaire par rapport à la signification intentionnelle, et un énoncé tel que *je voudrais savoir si le night-club appelé " le Guermante" se trouve bien à Martinville* devrait être interprété par un système comme une interrogation directe partielle (la réponse attendue ne peut pas se limiter par exemple à un simple «non»), alors que sa formulation est celle d'une interrogation indirecte totale. Il est clair ensuite que c'est ce night-club-là qui est recherché et non un autre, fût-il à Martinville, et c'est cette piste-là qu'il faudrait privilégier en cas de réponse négative.

Cette question du maniement des buts, c'est-à-dire de la signification intentionnelle a d'ailleurs été au centre de la quasi-totalité des recherches en intelligence artificielle relatives aux traitements automatiques du langage durant les années 1980-1990. Qu'on parle des **buts**, des **plans**, des **croyances**, des **foyers** ou des **univers de référence** (termes qui recouvrent des conceptualisations différentes) (cf. P. Cohen *& al.* 1990), on en revient toujours au problème de la signification intentionnelle : il est possible de considérer que l'on est parvenu à représenter les buts en machine (en ce sens qu'on sait les mettre

en rapport avec l'analyse ou même entre eux), on reste en revanche incapable de réaliser des systèmes qui conçoivent eux-mêmes de nouveaux buts. En d'autres termes, on en est au même point en ce qui concerne la signification intentionnelle qu'en ce qui concerne le sens propositionnel : un système doit nécessairement évoluer dans un univers quadrillé s'il veut prétendre à une quelconque efficacité. Or si cette nécessité est une contrainte naturelle en ce qui concerne l'analyse de phrase, dans la mesure où il peut paraître légitime que le système ait une représentation de tout ce qui est explicite (lexique et syntaxe notamment), c'est une contrainte beaucoup plus gênante pour ce qui est des buts. Cela revient en effet à tenter de maîtriser l'implicite sans avoir aucune prise sur l'imprévu, et cela condamne les tentatives de réalisation à se limiter à l'exécution de tâches suffisamment simples et délimitées pour qu'une "grammaire" des buts puisse servir de base à la réalisation informatique.

1.1.2. La non-normativité

A la différence de l'analyse, qui suppose très généralement une langue standard normée, le dialogue est nécessairement confronté à la non-normativité. Que ce soit à l'oral ou à l'écrit, le dialogue suppose une communication directe où la norme est d'emblée considérée comme secondaire. Dans les données recueillies avec machine simulée, on trouve ainsi, au-delà de quelques exemples caricaturaux (comme la question suivante, proposée par l'intermédiaire du Minitel, et destinée à un ministre de l'emploi : *il y aurait peut etre des problème de secret progessionnel???la ligeslation prevois t'elle se tavail???a qui je peux me renseigner???*), une langue qui comporte des phénomènes de **bruits** et des phénomènes de **dislocation** analogues à ceux que l'on rencontre usuellement à l'oral.

La langue en question n'est pas celle de l'écrit standard. Comme dans la langue parlée, le dialogue suppose que les énoncés soient conçus et perçus dans le fil de leur énonciation. Même avec un clavier, il est rare que l'on soit tenté de se corriger autrement que par un supplément de message, et on rencontre fréquemment divers phénomènes de non-normativité comme des télescopages syntaxiques tels que les suivants :

- Pour les banlieues il faut une communication beaucoup plus rapide de banlieue à banlieue.

- j'ai 15 ans je suis en troisième pourrais-je quitter l'école pour aller au CFA sans avoir 16 ans.

Ce type de non normativité relève moins d'une "grammaire des fautes" que d'un processus énonciatif particulier. La notion de cohérence syntaxique, au lieu de s'appliquer à un concept tel que celui de la phrase, s'applique à une fenêtre qui se déplace au fil de l'énonciation, et qui a pour effet de faire cohabiter des structures syntaxiques *a priori* exclusives les unes des autres : *pour les banlieues il faut une communication beaucoup plus rapide* et *il faut une communication beaucoup plus rapide de banlieue à banlieue* pour le premier

exemple ; *j'ai 15 ans je suis en troisième pourrais-je quitter l'école pour aller au CFA* et *je suis en troisième pourrais-je quitter l'école pour aller au CFA sans avoir 16 ans* pour le second.

Dans un certain nombre de cas (lorsque l'importance des phénomènes de non-normativité rend les techniques d'analyse classiques inopérantes notamment), le dialogue peut conduire au développement de techniques d'analyse spécifiques. Très souvent par exemple se pose le problème de l'analyse totale ou de l'analyse partielle : dès lors que les phénomènes de non-normativité sont importants, mieux vaut se limiter à analyser les passages dont on est sûr, et laisser de côté ceux qui sont susceptibles de faire difficulté. L'importance de la composante pragmatique pourra seule pallier dans ce cas les carences d'une analyse par définition lacunaire : cela ne fonctionne que si le système parvient, grâce à la maîtrise de la signification intentionnelle, à prévoir ce qui va être dit.

1.2. Dialogue et sens

Dans le dialogue, nous allons voir que le concept de sens doit être manipulé de façon particulière, à la fois parce que l'erreur n'a pas la même valeur en dialogue qu'ailleurs — dans la mesure où elle vaut davantage en fonction du coût de son éventuelle récupération qu'en elle-même — et parce qu'on peut discerner une instance de sens spécifique au dialogue, la signification interactionnelle. Celle-ci permet de rendre compte du jeu des questions et des réponses et, partant, de la structure du dialogue.

1.2.1. Erreur et dialogue

Le dialogue a toutefois une spécificité : son caractère dynamique, qui s'oppose au caractère statique de l'analyse. Cette **dynamicité** s'illustre avec profit à partir du concept d'erreur. En situation d'analyse, l'erreur se juge par rapport à une analyse théoriquement irréprochable admise une fois pour toutes, et elle se jauge à l'aide de paramètres divers, en fonction de la distance qui sépare l'analyse de référence et l'analyse effectuée par le système. En situation de dialogue, l'erreur ne se juge pas en elle-même. Elle se juge en fonction du coût éventuel de sa récupération.

On peut ainsi distinguer différents types d'erreur, ne serait-ce qu'à partir d'un exemple tel que la demande d'horaire SNCF suivante :

- j'aimerais me rendre à Lorient de Paris tard ce soir et je voudrais savoir dans quelle gare je devais partir.

Les erreurs **négligeables** sont celles qui n'influencent pas le déroulement du dialogue. Ainsi, même si la partie de la requête portant sur la nature de la gare parisienne n'était pas prise en compte, n'importe quelle réponse inclurait cette information, ce qui retirerait toute pertinence à l'erreur en question.

Les erreurs à **rectification immédiate** sont celles qui ne nécessitent qu'une demande de précision ou de rectification, de la part de la machine.

Ainsi, au cas où l'indication du jour, par l'intermédiaire du démonstratif *ce*, n'était pas comprise, cela supposerait la formulation immédiate d'une question incidente (*Quel jour désirez-vous partir ?*) de la part du système.

Les erreurs à **rectification différée** sont celles qui supposent que la demande de précision ou de rectification émane du correspondant, à la suite d'une réponse partiellement fausse de la machine présumée. Ainsi, dans l'hypothèse où *tard* resterait incompris, le système proposerait une réponse inadéquate, ce qui conduirait l'interlocuteur à solliciter une nouvelle réponse.

Les erreurs **non-rectifiables** sont celles qui nécessitent une reprise de toute la séquence de dialogue. Il s'agit en fait des erreurs qui supposent une réelle incompréhension de la requête, comme si le système comprenait que la question portait sur le trajet Lorient-Paris par exemple : une réponse peut être proposée et le dialogue s'engager sans que l'interlocuteur ne se rende compte de la méprise (d'où l'intérêt bien sûr des fréquents échanges confirmatifs en dialogue homme-machine ; cf. E. Bilange 1992).

1.2.2. La signification interactionnelle

Les erreurs sont en tout cas moins importantes en dialogue que le maintien du canal de communication, qui est indispensable à la récupération éventuelle de ces erreurs. Il existe en somme une forme particulière de sens qui relève de ce que l'on appellera la **signification interactionnelle.** Considérons en l'occurrence l'exemple suivant :

H1 : c'est la météo nationale ?
M1 : *oui*
H2 : quel temps fera-t-il demain ?
M2 : *où cela ?*
H3 : connaissez-vous Martinville ?
M3 : *non*
H4 : c'est à côté de Tansonville
M4 : *oui*
H5 : alors de ce côté-là
M5 : *il y fera un temps superbe*
H6 : est-ce que les aubépines seront en fleurs ?
M6 : *je ne sais pas*
H7 : pourquoi ?
M7 : *les problèmes de temps perdu ne me concernent pas*
H8 : tant pis, il fera beau de toute façon, n'est-ce pas ?
M8 : *oui.*

La séquence H6-M8 ne pose, du point de vue dialogique, aucun problème, même si on considère qu'il est très peu probable qu'un système réponde un jour comme en M7. En revanche H2, H3, H4 et H5 visent à satisfaire un seul et même but : savoir le temps qu'il fera "du côté de...", mais tantôt la syntaxe est interrogative et tantôt elle ne l'est pas ; parfois l'interrogation est totale

(H3), parfois elle est partielle (H2). Signification intentionnelle et sens propositionnel ne rendent ainsi absolument pas compte de cette forme particulière de sens qui fait que, pour des raisons proprement dialogiques :

- H3 est davantage une réponse à M2 qu'une quelconque question, et s'il est tout à fait acceptable de feindre, comme en M3, que c'est une question, c'est parce que Martinville est inconnue et que cela évite toute forme de réitération de M2. Au cas où Martinville serait connue, le système devrait en revanche être capable de fournir immédiatement en M3 la réponse qui intervient en M5. C'est d'ailleurs l'absence d'un tel comportement dialogique qui rend M4 difficilement acceptable : en toute logique, c'est à ce moment-là que le système devrait répondre.

- La maladresse que dénote M4 constitue une erreur négligeable, dans la mesure où cela ne gêne en aucune façon la poursuite du dialogue. Elle tient à une mauvaise interprétation de H4, compris, en rapport avec H3, comme une paraphrase de **connaissez-vous Tansonvillle*, alors que cette intervention devrait être interprété comme une réponse à M2. Nous nous interrogerons toutefois sur les incidences de cette erreur par la suite (cf. *infra,* § 2.2.2.).

- Indépendamment des mécanismes d'inférence nécessaires à son analyse, H5 est, du point de vue de la signification interactionnelle, une réponse à M2, même si cela peut être ressenti comme une paraphrase de **quel temps fera-t-il demain du côté de Tansonville*, qui interviendrait comme une relance à la suite d'une intervention maladroite du système.

1.3. Problème des tâches

Comme on le voit, tout en se limitant à des tâches opératives (c'est-à-dire qui visent à la réalisation d'actions simples), le dialogue homme-machine pose des problèmes dialogiques complexes. Il souffre par ailleurs de divers maux qui perturbent son appréhension : le dialogue homme-machine est ainsi rarement considéré comme une fin en soi (c'est essentiellement un faire-valoir de performances qui, bien souvent, sont même étrangères aux traitements automatiques du langage : cf. A. Vilnat, 1984) ; le dialogue homme-machine est ensuite spontanément identifié au dialogue homme-homme, ce dernier, au demeurant assez mal connu et relativement peu étudié, constituant implicitement le cadre mental de référence. Au total donc, le domaine du dialogue homme-machine reste restreint, que ce soit parce que dans les faits nous adoptons un comportement langagier induit par la machine, ou parce que la nature des tâches reste fatalement extrêmement limitée.

1.3.1. L'interlocuteur machine

Il suffit pourtant, pour se convaincre du contraire, d'observer comment nous nous exprimons lorsque nous nous trouvons placés en situation de communication avec une machine, ce que l'on appelle le **comportement langa-**

gier induit par la machine. Aussi bien en ce qui concerne la construction des énoncés qu'en ce qui concerne la gestion du dialogue, on aboutit à des attitudes langagières tout à fait spécifiques.

Face à une opératrice, un locuteur demandera par exemple *oui bonjour madame je voudrais connaître les trains Paris Le Mans ce soir à partir de 6 heures* ; face à une machine simulée, ce même locuteur dira *horaire des trains Paris-Montparnasse Le Mans le vendredi 28 décembre à partir de 18 heures,* ce qui dénote, au-delà d'un style télégraphique et d'une dépersonnalisation, une **évolution du repérage spatio-temporel** qui rend une éventuelle analyse automatique bien plus confortable.

Pour ce qui est du dialogue, on passe d'un dialogue à **déroulement linéaire**, dans lequel la cohérence se limite souvent aux échanges consécutifs, à un dialogue à **structure hiérarchique**, dans lequel l'ensemble du dialogue obéit à une structure rigide. Dans un cas, on passera aisément du coq à l'âne, pour revenir éventuellement aux gallinacés, alors que dans l'autre, il n'est pas question, dans l'esprit des interlocuteurs, de changer de sujet sans avoir conclu quant au précédent, fût-ce par un constat d'échec (pour de plus amples détails, cf. M.A. Morel *& al.* 1988-89, ou le n° 509 du *Monde informatique* du 13.07.92 : "Homme-machine : un dialogue à réinventer").

Face à une machine, la communication verbale devient en somme spontanément plus rigide, comme si cette rigidité était actuellement ressentie comme nécessaire à la fois pour permettre au système de suivre, et pour garder ses distances face à un interlocuteur perçu comme non-humain (nul ne peut savoir ce qu'il en sera, à l'usage d'une part, et lorsque les systèmes auront évolué d'autre part). Dialoguer avec une machine et dialoguer avec un interlocuteur humain sont en tout cas ressentis comme des processus radicalement différents. Le dialogue homme-machine est un type nouveau de communication, qui ne fait que poindre, qui évoluera tout autant en fonction de la perception que nous en aurons que de ses performances techniques, qu'il nous reste en grande partie à inventer.

1.3.2. Le dialogue opératif

Pour un grand nombre d'informaticiens (en l'occurrence de réalisateurs de systèmes de dialogue), le dialogue est moins une fin en soi qu'un faire-valoir de performances diverses. Ainsi, deux des principaux systèmes de dialogue des années 70 peuvent-ils être considérés comme des systèmes dans lesquels le dialogue est secondaire :

- GUS (*Genial Understanding System*) (cf. D. Bobrow *& al.* 1977), censé permettre la réservation de places d'avion, est tout autant destiné au développement de KRL (nouveau langage de programmation conçu pour faciliter la réalisation de tels systèmes) qu'à celui d'un système de dialogue homme-machine en lui-même.

- SHRDLU (ainsi nommé pour la difficulté de la prononciation de ce symbole de la coquille dans certains magazines américains) (cf. T. Winograd 1983), consacré à la manipulation de formes géométriques sur écran, est avant tout un système de compréhension automatique du langage naturel qui illustre le traitement de pronoms déictiques ou représentants (*find a block which is taller than the one you are holding and put it onto the box*), et dans lequel le dialogue sert essentiellement à illustrer le fonctionnement de la compréhension.

Plus récemment, un certain nombre de systèmes de dialogue homme-machine ont été réalisés, dans lesquels c'est le fonctionnement du dialogue qui est visé. Lorsque ce sont des systèmes avec entrée au clavier, l'alternative est simple : soit la tâche est extrêmement limitée, et le système fonctionne (par exemple les renseignements horaires SNCF avec des contraintes liées aux trajets) ; soit la tâche est plus vaste et plus intéressante, et le système est un prototype de laboratoire qui valide certaines hypothèses mais qui ne prétend à aucune efficacité. Lorsque ce sont des systèmes avec entrée vocale, les contraintes sont encore plus fortes, dans la mesure où on n'a plus affaire qu'à un dialogue de commande extrêmement figé (cf. M. Guyomard *& al.* 1990).

2. Modélisations et réalisations informatiques

Aussi bien pour les linguistes que pour les informaticiens, le dialogue verbal reste, comme on va le voir, un domaine de recherche relativement marginal. Les travaux réalisés dans ce domaine ont toutefois permis l'émergence de modélisations informatiques : après avoir présenté un modèle hiérarchique, puis un modèle dynamique, on examinera les composantes classiques des systèmes actuels, qui comportent un même ensemble de connaissances, les unes statiques, les autres dynamiques.

2.1. L'étude du dialogue

Tant du côté des linguistes que des informaticiens, le dialogue verbal a jusqu'ici donné lieu à un nombre limité de travaux, tous de date récente. Ce sont en outre des approches qui ont du mal à collaborer entre elles, du fait des différences de préoccupations entre les deux disciplines.

2.1.1. Différentes approches

En linguistique, l'étude du dialogue est souvent ressentie comme une application de la pragmatique davantage que comme un domaine autonome. L'étude du dialogue en tant que tel est d'ailleurs récente et, en dépit de divers travaux très intéressants, elle se limite pour nous à deux courants majeurs : les conversationnalistes américains et l'école de Genève.

Les **conversationnalistes américains** (cf. H. Sacks *& al.* 1974) : il s'agit avant tout de sociologues, dont la caractéristique première est de s'être attachés à une description **informelle** de la conversation, sans se soucier de bâtir des systèmes explicatifs. Ils ont ainsi mis en évidence le rôle déterminant des rapports psychologiques entre les intervenants, et l'importance des gestes, silences, interruptions, reprises et autres "bruits" dans la conversation.

L'**école de Genève** (cf. E. Roulet *& al.* 1985) : il s'agit, tout en partant également de conversations réelles, de bâtir un système descriptif aussi structuré que possible. Il en ressort un modèle de description fondé à la fois sur le courant conversationnaliste (le dialogue est conçu comme une négociation), sur la pragmatique anglo-saxonne, et sur la théorie de l'argumentation. L'**intervention**, qui ne s'identifie pas avec les tours de parole, comporte différents actes de langage hiérarchisés. Elle s'intègre dans un échange, car une intervention initiative appelle une intervention réactive. Deux contraintes, issues des maximes de Grice (cf. H.P. Grice 1979), déterminent dans une large mesure la structure des discours en situation interactive : la **complétude interactionnelle** (tendance à faire progresser le dialogue vers la satisfaction des deux intervenants), et la **complétude interactive** (tendance, lorsqu'il y a désaccord, à la résolution des conflits pour pouvoir revenir au dialogue principal et envisager de satisfaire la complétude interactionnelle).

D'autres domaines, et notamment l'**ergonomie**, se sont également, en marge du dialogue homme-machine, intéressés à l'étude du dialogue (cf. P. Falzon, 1989). Il s'agit en l'occurrence d'étudier les conditions de réalisation et de pertinence de systèmes de dialogue homme-machine, ainsi que le contexte des études linguistiques sur le sujet : dans quelle mesure dialogue homme-homme et dialogue homme-machine sont-ils similaires ? La méthode dite du "Magicien d'Oz" (avec machine simulée) permet-elle d'obtenir des résultats fiables ? Celui qui simule la machine peut-il éviter de conditionner en grande partie le comportement des usagers ? Qu'en sera-t-il lorsque ces derniers se seront habitués à ce type de communication ?

Dans ce domaine du dialogue, plus peut-être encore que dans d'autres, on peut constater des divergences considérables entre l'Europe et les États-Unis : la seule théorie descriptive qui soit à notre sens suffisamment rigoureuse pour être susceptible de servir à la réalisation de programmes de dialogue homme-machine est européenne et francophone. Alors qu'elle sert de référence dans l'ensemble des équipes françaises travaillant sur le sujet, elle est ainsi, depuis presque dix ans, à peu près totalement ignorée dans l'ensemble des pays à tradition anglo-saxonne.

2.1.2. Linguistique et informatique

Le domaine du dialogue homme-machine offre un bon exemple des difficultés que peut rencontrer la collaboration entre informatique et linguistique, du fait du statut même de ces deux disciplines : la linguistique vise à décrire

la richesse de la langue, alors que les réalisations informatiques tendent inéluctablement à une simplification réductrice. Deux exemples sont à cet égard éclairants : d'une part les modélisations de P. Cohen *& al.* (1990), ou de J. Allen & C.R. Perrault (1980) à partir de J-L. Austin et de J. Searle, qui font partie des contributions les plus marquantes des dix dernières années en dialogue homme-machine ; d'autre part les efforts de modélisation de E. Roulet *& al.* (1985) et de J. Moeschler (1985) et (1989), qui ont servi de base à presque tous les plus récents systèmes de dialogue homme-machine en France.

Pour rendre compte de la signification intentionnelle, J. Allen & C.R. Perrault se sont efforcés d'utiliser la théorie des actes de langage pour traiter les demandes de renseignements suivantes, en partant du constat que, comme pour les exemples sur le sel de J. Searle (*avez-vous du sel ? pouvez-vous me passer le sel ?*) une réponse coopérative devrait tenir compte non seulement de la fonction locutoire, mais également des fonctions illocutoires et perlocutoires :

- *Quand part le prochain train pour XXX*
- *Prochain train pour XXX*
- *Savez-vous quand part le prochain train pour XXX*
- *Est-ce que le prochain train pour XXX part*
- *Savez-vous si le prochain train pour XXX part*
- *Sauriez-vous si le prochain train pour XXX part.*

Le linguiste constatera qu'un arsenal théorique complexe n'est pas indispensable, et qu'une analyse tout à fait pertinente est réalisable à partir de la syntaxe de l'interrogation : l'ensemble de ces formulations est assimilable à une interrogation directe partielle (la réponse attendue comporte toujours l'horaire du train en question), c'est-à-dire au premier exemple, et les différentes formulations peuvent simplement induire un éventuel "habillage" différent de la réponse (*15h quai 2 / le prochain train pour XXX part à 15h du quai 2*). Ce n'est en définitive qu'au cas où il n'y aurait aucun train que les réponses pourraient éventuellement diverger. Cela dit, cette utilisation de la théorie des actes de langage a conduit au développement de concepts aujourd'hui couramment utilisés, tels que les concepts de **buts**, ce que cherche un locuteur en posant une question (se rendre à XXX en demandant l'heure du train qui y va par exemple), de **plans**, ce que fait le locuteur pour réaliser son but (demander l'heure du train pour XXX quand on veut s'y rendre par exemple) ou d'**obstacles**, ce qui ferait défaut pour réaliser un but (le quai d'où part le train recherché par exemple).

Développé pour aider à la description de conversations réelles, le modèle genevois, quant à lui, vise à rendre compte de dialogues complexes, dans lesquels chaque tour de parole comporte souvent plusieurs actes de langage. En dialogue homme-machine, le type d'interaction diffère, et il est bien rare qu'un tour de parole excède un acte de langage. Tout au plus rencontre-t-on

des rituels d'évaluation, souvent appelés échanges confirmatifs (cf. H. Sacks *& al.* 1974), qui consistent à approuver ou à désapprouver, à donner son accord ou à signaler son désaccord, à manifester sa compréhension ou son incompréhension, parfois par une reprise (*Je voudrais un train Paris Lyon / Paris Lyon, à quelle heure*), parfois par une marque linguistique (cf. H8 : *tant pis*, dans l'exemple sur la météo repris ci-dessous).

Ce modèle, qui permet de produire une représentation arborescente (cf. *infra*, § 2.2.1.), est fondamentalement un modèle descriptif, qui permet de rendre compte d'une conversation une fois celle-ci achevée. Le problème en dialogue homme-machine est à l'inverse de rendre compte d'un dialogue au cours de son déroulement, de façon **dynamique**, afin de pouvoir autant que possible en garder le contrôle. Cela dit, un peu comme les théories de J-L. Austin et de J. Searle, le modèle genevois a rendu des services considérables. A notre sens, il a notamment permis de dégager que tout dialogue ne pouvait s'orienter que dans deux directions complémentaires : on a affaire soit à un **dialogue régissant,** soit à un **dialogue incident** (cf. *infra*, § 2.2.2.). Le premier vise à réaliser les buts successifs des locuteurs ; le second résout les difficultés ponctuelles qui gênent le déroulement du premier (incompréhension ; manque d'information...). Ainsi, dans l'exemple ci-dessous, les séquences M2-H5 et H7-M7 sont-elles des séquences incidentes, alors que tout le reste relève du dialogue régissant.

2.2. Modélisations du dialogue

Avant de décrire le fonctionnement global des systèmes de dialogue, nous présenterons brièvement certaines des modélisations qui en constituent le centre. Il s'agit en l'occurrence de diverses façons de représenter le dialogue en machine.

2.2.1. Un modèle hiérarchique

Le modèle genevois a induit, dans les travaux sur le dialogue homme-machine, des représentations hiérarchiques telles que celle présentée dans la Figure II.

Ce type de représentation, au départ fondée sur l'utilisation des concepts d'acte de langage, d'intervention et d'échange, peut être complété par l'étiquetage des branches à partir de la fonction illocutoire des tours de parole correspondants, mais elle est avant tout intéressante dans un système par le mécanisme d'hiérarchisation qui est mis en oeuvre. Chaque ligne correspond, sinon à une intervention, du moins à un acte de langage : dans notre exemple, seul H8 comporte deux actes de langage, ce qui donne lieu à deux interventions qui s'intègrent dans des échanges différents (H8 et H8'). Lorsque les couples de question-réponse sont contigus, cela donne lieu à des échanges simples (H1-M1, H3-M3, H4-M4, H8'-M8), parfois à trois composantes (H7-M7-H8) ; la réponse, comme c'est le cas à la suite de M6, peut inclure une séquence explicative. Lorsque les couples de question-réponse sont disconti-

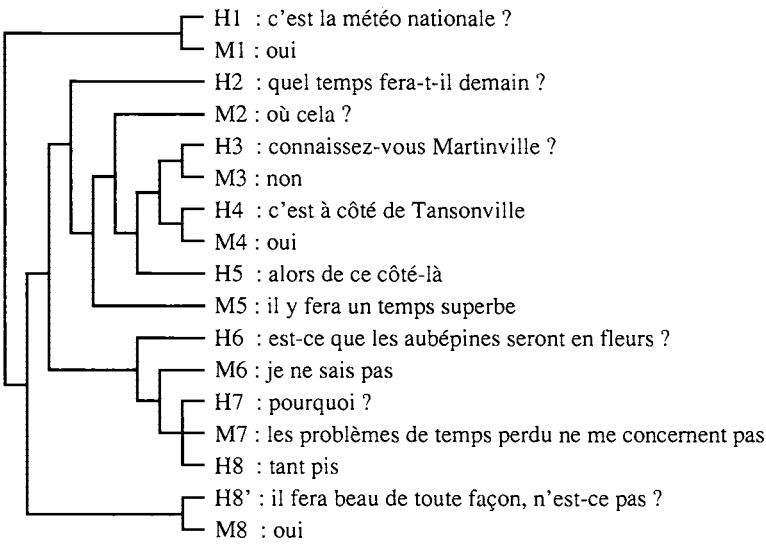


FIGURE II

nus, on a alors affaire à des échanges complexes : M5 est la réponse à H2, et cette réponse inclut la séquence M2-H5 ; à l'intérieur de cette séquence, H5 est la réponse à M2, et cette réponse inclut la séquence H3-M4.

Ce type de procédé est tout à fait comparable, en syntaxe, avec la représentation arborescente d'une phrase : c'est avant tout une méthode de représentation et beaucoup plus difficilement une méthode calculatoire ; cela ne fonctionne efficacement que sur une séquence à la fois bien formée et achevée. On peut d'ailleurs recourir à d'autres représentations similaires, que ce soit sous forme de formule parenthésée :

((H1 M1)(H2(M2(H3 M3)(H4 M4)M5)(H6 M6(H7 M7 H8))(H8' M8)))

ou sous une forme comparable à une boîte de Hockett :

				H3	M3	H4	M4									
				H3	M3	H4	M4					H7	M7	H8		
			M2	H3	M3	H4	M4	H5			M6	H7	M7	H8		
		H2	M2	H3	M3	H4	M4	H5	M5	H6	M6	H7	M7	H8		
		H2	M2	H3	M3	H4	M4	H5	M5	H6	M6	H7	M7	H8	H8'	M8
H1	M1	H2	M2	H3	M3	H4	M4	H5	M5	H6	M6	H7	M7	H8	H8'	M8
H1	M1	H2	M2	H3	M3	H4	M4	H5	M5	H6	M6	H7	M7	H8	H8'	M8

2.2.2. Un modèle dynamique

En s'inspirant du modèle hiérarchique, on peut construire un modèle dynamique (cf. D. Luzzati 1989), ce qui permet, par une représentation calculatoire de la structure du dialogue, de gérer celui-ci dynamiquement, c'est-à-dire de réagir en cours de communication (voir Figure III).

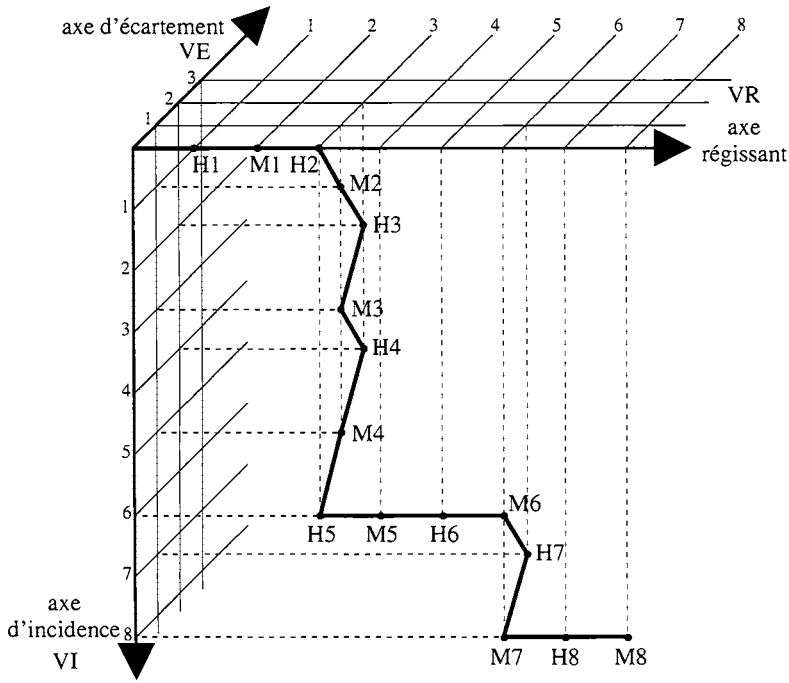


FIGURE III

Le dialogue peut s'orienter dans deux directions : soit demande d'information et délivrance des renseignements s'enchaînent sans difficulté, et il s'agit d'un dialogue **régissant** ; soit des demandes de précision, d'explication, de confirmation ou de reformulation doivent intervenir pour qu'une question ou une réponse soit acceptée, et il s'agit d'un dialogue **incident**. Ainsi, à la suite de H2, le système ne peut-il pas répondre sans disposer de références spatiales, tout comme à la suite de M6, le correspondant souhaite voir justifiée la réponse qui vient de lui être fournie : on assiste alors à l'ouverture de deux axes incidents, le premier à l'initiative de la machine, le second à l'initiative du correspondant. Le dialogue suit nécessairement soit l'axe régissant, soit l'axe incident, VR et VI s'incrémentant de façon exclusive. Le troisième axe quant à lui n'intervient que dans le courant d'un axe incident, en indiquant la distance par rapport au retour potentiel sur l'axe

régissant : VE s'incrémente lorsqu'il s'agit de questions incidentes (M2, H3, H4, H7), et se désincrémente lorsqu'il s'agit des réponses correspondantes (M3, M4, H5, M7).

Ce qui est important, en matière de modèle de dialogue, c'est moins la représentation en elle-même que les décisions qu'elle permet de prendre dans le courant de la communication. Ainsi, dans l'exemple proposé, pourrait-il être préférable que H3 et H4 puissent être comprises comme des réponses, au lieu d'être interprétées comme des questions, la réponse à H2, sur l'axe régissant, intervenant alors en M3 ou en M4. C'est précisément ce que permet le modèle, à partir de la fonction suivante :

$$F=(x*VR)+(y*VI)+(z*VE)$$

dans laquelle x, y et z peuvent être prédéfinis, ou bien varier dynamiquement, en fonction des modèles de la tâche (cf. *infra*, 2.3.1.) et de l'utilisateur (cf. *infra*, § 2.3.2.) par exemple. Pour ce qui concerne l'interprétation de H3 et de H4, en attribuant respectivement les valeurs de 2, 5 et 20 à x, y et z (ce qui confère un poids négligeable au dialogue régissant, qui ne pose aucun problème, et un poids d'autant plus considérable au dialogue incident qu'il s'écarte du niveau de retour sur l'axe régissant), on obtient :

$$F(H3)=56$$
$$F(H4)=96$$

alors que si ces mêmes interventions étaient interprétées comme des réponses, on aurait :

$$F(H3)=16$$
$$F(H4)=26.$$

Selon les cas, la valeur la plus faible de F peut être systématiquement privilégiée, ce qui revient à refuser l'incidence et à figer l'interaction, ou bien on peut accepter l'incrémentation de VE dès lors que F ne dépasse pas un certain seuil.

2.3. Les systèmes de dialogue homme-machine

Un tel modèle ne constitue pas à lui seul un système de dialogue homme-machine, même si c'est à lui que revient la responsabilité de gérer le dialogue proprement dit. Il représente seulement quelques-unes des connaissances que la machine doit posséder, dont certaines sont statiques, c'est-à-dire qu'elles ne varient pas au cours du dialogue, et certaines sont dynamiques, dans la mesure où elles se modifient dans le cours de la communication. Ces connaissances, telles qu'elles sont représentées dans la Figure IV ci-dessous, ne constituent pas une architecture opérationnelle. Tout système de dialogue homme-machine doit les intégrer mais, d'une part, elles n'y ont pas toujours la même importance, et d'autre part elles sont trop interdépendantes pour pouvoir se répartir en modules distincts, comme pourrait le laisser supposer une présentation schématique telle que celle de la Figure IV.

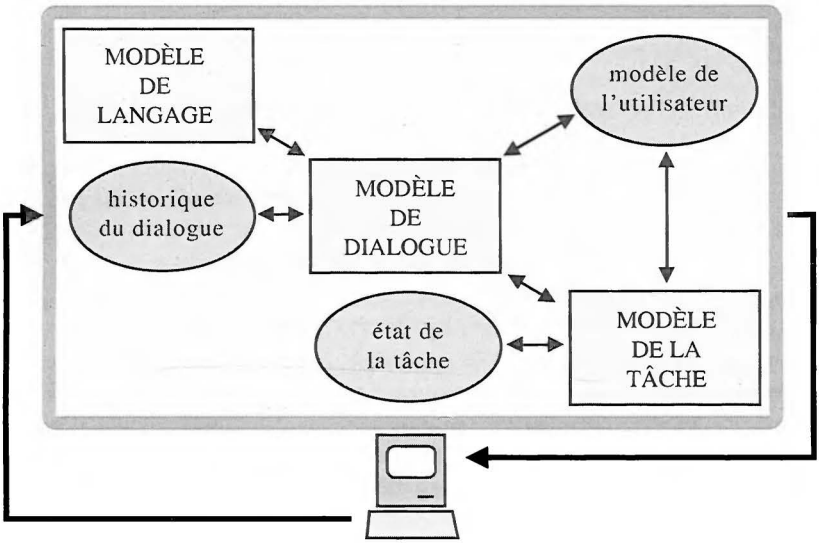


FIGURE IV

2.3.1. Connaissances statiques

On a coutume de dissocier en trois parties les connaissances initiales dont doit disposer un système de dialogue homme-machine : le modèle de langage, le modèle de la tâche et le modèle de dialogue.

Le **modèle de langage**, qui comporte les connaissances linguistiques nécessaires au fonctionnement de l'analyseur, est souvent considéré comme extérieur au fonctionnement du dialogue proprement dit. Les systèmes de dialogue homme-machine requièrent toutefois fréquemment des connaissances lexicales, syntaxiques et sémantiques adaptées à une tâche, c'est-à-dire par certains aspects très détaillés, et par d'autres extrêmement limités. Dans un système de renseignements horaires SNCF par exemple, *train* a tout intérêt à être systématiquement compris comme «place assise» (*je voudrais un train pour Paris demain matin*), en non comme «ensemble de wagons tirés par une locomotive», avec toutes les valeurs d'emplois que cela peut comporter.

Le **modèle de la tâche** a pour fonction de représenter et de structurer l'univers de référence, dans lequel est censé évoluer le système. Ainsi, dans l'exemple proposé, est-il éminemment souhaitable, avant même qu'on lui ait parlé de *Martinville* ou de *Tansonville*, que le système se soit attendu à se voir proposer des toponymes, de la même façon qu'il n'a aucune chance d'interpréter correctement *aubépines* s'il n'est pas prévu que l'on évoque les fleurs à propos de renseignements météo. Le modèle de la tâche a donc pour

fonction essentielle d'assurer la gestion thématique du dialogue. Il est également indispensable pour lever les ambiguïtés liées aux ellipses et aux anaphores, particulièrement fréquentes en dialogue : des interventions comme *le suivant - celui d'avant - plus tôt - plus tard* sont parfois omniprésentes. Cette représentation de l'univers de référence peut notamment se faire par le biais de techniques de représentation évoquées au chapitre 8, telles que schémas, scénarios, plans ou Mops, et elle se gère par l'intermédiaire de l'historique du dialogue.

Le **modèle de dialogue**, sur un exemple duquel nous nous sommes attardés (cf. *supra*, § 2.2.), constitue la partie centrale d'un système de dialogue homme-machine. A la différence des deux précédentes parties, qui peuvent concerner toute forme de discours, celle-ci est exclusivement réservée au dialogue proprement dit. Sous une forme thématique (classification des échanges en échanges informatifs, confirmatifs, évaluatifs...), ou sous une forme hiérarchique (on pourrait presque dire "syntaxique"), telle que nous l'avons décrite, le modèle doit présenter une "grammaire" du dialogue qui permette d'identifier et de classer les différentes séquences d'interventions.

2.3.2. Connaissances dynamiques

On subdivise également en trois parties les connaissances qui évoluent au cours d'une communication : on distingue ainsi l'état de la tâche, l'historique du dialogue et le modèle de l'utilisateur.

L'**état de la tâche**, associé au modèle de la tâche, a pour fonction d'accorder le niveau de compréhension du système avec l'état d'avancement de la tâche. *Direct* par exemple peut selon les cas signifier «sans arrêt» ou bien «sans changement» ; *le premier* s'interprétera comme une date si l'intervention qui précède est la question *quel jour désirez-vous partir ?*, comme une abréviation de «premier train» s'il suit la question *à quelle heure désirez-vous partir ?*, et comme «le premier train que vous m'avez proposé» après la question *lequel de ces trains vous convient-il ?* Dans un cas comme dans l'autre, ce sera l'état de la tâche qui permettra de lever l'ambiguïté.

L'**historique du dialogue**, associé au modèle du dialogue, a pour fonction de conserver la structure et le contenu du dialogue en cours. Une demande de reformulation par exemple ne peut pas intervenir de la même manière au début d'un dialogue ou à la suite de plusieurs échanges improductifs. De la même manière, il est nécessaire de gérer les changements de thème car, comme dans l'exemple analysé ci-dessus, on peut revenir à un thème antérieur (le temps qu'il fera à Martinville) après avoir fait une digression (l'épanouissement des aubépines). Pour ces questions de contenu, on peut utiliser les modélisations de P. Cohen, J. Allen & C.R. Perrault, en termes de buts et de plans, alors que pour ce qui est de la structure du dialogue, c'est une des fonctions du modèle hiérarchique présenté ci-dessus, auquel on peut associer un agenda des interventions.

Le **modèle de l'utilisateur**, associé à la fois au modèle de dialogue et à celui de la tâche, doit permettre au système de s'adapter à son interlocuteur, ce qui est fondamental dans certaines applications, et relativement secondaire dans d'autres. L'utilisateur d'un système de renseignements horaires SNCF par exemple peut être considéré comme un utilisateur type, *a priori* expert dans le domaine (il sait ce qu'est une gare, un train, une place, une réservation...), et ce n'est qu'avec des situations exceptionnelles qu'une représentation de l'interlocuteur devient importante (problèmes de congés payés, transports d'animaux, renseignements pris pour un tiers...). Dans les systèmes d'enseignement intelligemment assistés par ordinateur (EIAO) ou pour des tâches plus complexes, la représentation de l'interlocuteur devient l'élément central du système. Ainsi, dans l'exemple "pages jaunes" proposé plus haut (cf. *supra*, § 1.1.1.), a-t-on un correspondant qui cherche un garage et qui demande ensuite un hôtel. Cela suppose une succession d'inférences complexes dans la représentation que le système possède de cet interlocuteur : il doit faire réparer sa voiture ; la réparation peut être longue ; il va être à pied ; l'hôtel et le garage doivent être proches ; peut-être souhaitera-t-il qu'on lui propose des occupations pendant son inactivité forcée... Ici également, c'est le plus souvent une représentation en termes de buts et de plans qui est la plus souvent retenue.

3. Perspectives

Les systèmes de dialogue homme-machine de référence, GUS et SHDRLU notamment (cf. *supra*, § 1.3.2.) datent des années 70, et on peut légitimement se demander dans quelle mesure on a réellement progressé au cours des vingt dernières années. La réponse des spécialistes sera certainement positive, dans la mesure où la modélisation de la signification intentionnelle et de la signification interactionnelle est récente et tout à fait prometteuse ; celle des utilisateurs vraisemblablement beaucoup plus tempérée, car on ne peut pas dire que le marché soit envahi par des systèmes performants et appréciés, alors même que le Minitel offre des possibilités de diffusion considérables.

La **fracture** entre chercheurs et utilisateurs existe d'ailleurs depuis longtemps, et elle revêt une tonalité particulière en dialogue homme-machine. Le système qui a connu la plus grande notoriété, ELIZA (cf. J. Weizenbaum 1976), était en effet une supercherie : il s'agissait de proposer une psychanalyse de type rogérienne (qui renvoie ses propres paroles au patient) ; la presse s'en empara et l'Amérique entière y crut, alors même que le système en question, se limitant à manipuler des énoncés à trous, ne comprenait strictement rien et que son auteur s'évertuait à le clamer. ELIZA est ainsi révélateur du fossé qui sépare profanes et spécialistes : les uns jugent simplement sur la

fluidité de l'interaction, alors que les autres s'intéressent essentiellement à la manipulation du sens que l'interaction dénote. Un système astucieux pourra ainsi aisément séduire les uns et laisser les autres indifférents.

La diffusion de systèmes de dialogue homme-machine repose en définitive sur le problème des **tâches**. Avec entrée au clavier, on devrait voir rapidement se développer des systèmes effectuant des tâches simples telles que la consultation de bases de données (renseignements SNCF, catalogue de vente par correspondance, fichiers divers...) : tout en apportant un confort d'utilisation certain, cela ne présente aucun danger pour les bases de données en question. Le problème devient plus délicat dès lors qu'il y action sur les bases de données (réservations SNCF, commandes à distance, modifications de fichiers...), car les erreurs peuvent porter bien davantage à conséquence. Quant aux tâches plus ambitieuses, qui nécessitent quasiment une représentation du monde pour être efficaces, elles continueront encore longtemps à animer les débats de laboratoire, en dépit de réalisations qui pourront être astucieuses et profitables. Avec entrée vocale, tout est bien entendu lié aux progrès de la reconnaissance de la parole multilocuteur par mots enchaînés, et il n'est pas impossible qu'à moyen terme on soit capable de faire avec entrée vocale ce que l'on fait actuellement avec entrée au clavier.

Divers enjeux **économiques** peuvent également influencer le développement des systèmes de dialogue homme-machine. Ceux-ci sont ainsi susceptibles de devenir l'élément-clé d'une politique commerciale : par leur intermédiaire, on peut notamment orienter une clientèle, qui vers certains vols ou certains trains, qui vers certains articles. Pour de nombreuses entreprises les outils télématiques sont en outre devenus à ce point vitaux que toute évolution les concernant (l'introduction de dialogue en langue naturelle en l'occurrence) peut mettre en cause la vie de l'entreprise. De plus, le volume des services concernés est tellement colossal que tout alourdissemnt des programmes doit être maîtrisé : en 1991, la SNCF par exemple a enregistré 22 millions de communications via le Minitel (information communiquée par la Direction de la Recherche de la SNCF), et la troisième société de vente par correspondance française -la CAMIF- 2,7 millions. Ce qui est enfin l'objet de nombreuses études, c'est le **dialogue multi-modal**, c'est-à-dire le dialogue par divers **média** : langage écrit et oral bien sûr, mais aussi vision et geste, par l'intermédiaire d'une grande variété d'outils : souris (en deux ou trois dimensions), écran tactile, crayon de désignation, vêtements numériques digitaux (gant notamment), casque de visualisation stéréoscopique... ; langage, image et geste sont ainsi mis en concurrence en fonction de leur efficacité dans la communication entre l'homme et la machine.

Daniel LUZZATI
(Université du Maine et LIMSI-CNRS)

Repères bibliographiques

1. Ouvrages généraux :

[Le livre de E. Roulet *& al.* présente le modèle genevois, et ceux de J. Moeschler en développent la modélisation. Les ouvrages de T. Winograd et de G. Sabah sont des sommes de référence qui comportent quelques chapitres sur le dialogue. Celui de J. Weizenbaum est une réflexion sur le dialogue homme-machine à partir d'ELIZA. Ceux de M-A. Morel *& al.* développent l'analyse du comportement langagier induit par la machine, et le livre de J-M. Pierrel fait le point sur le fonctionnement et sur les perspectives en matière de dialogue oral. Quant à l'ouvrage de E. Bilange, il propose un survol très complet des recherches sur la communication homme-machine en langue naturelle.]

BILANGE, E. (1992) : *Le dialogue naturel avec une machine*, Paris, Hermès.

CARRE, R. *& al.* (1991) : *Langage humain et machine*, Paris, Presses du CNRS.

[voir le ch. 5 : "Les mécanismes de dialogue".]

MOESCHLER, J. (1985) : *Argumentation et conversation. Eléments pour une analyse pragmatique du discours,* Paris, Hatier-Credif.

MOESCHLER, J. (1989) : *Modélisation du dialogue. Représentation de l'inférence argumentative*, Paris, Hermès.

MOREL, M.A. *& al.* (1988-1989) : *Analyse linguistique d'un corpus de dialogue homme-machine*, Tomes 1 et 2, Publications de la Sorbonne nouvelle.

PIERREL, J.M. (1987) : *Dialogue oral homme-machine*, Paris, Hermès.

ROULET, E. *& al.* (1985): *L'articulation du discours en français contemporain*, Berne, Lang.

SABAH, G. (1988) et (1989) : *L'intelligence artificielle et le langage*, 2 volumes, Paris, Hermès.

[sur le dialogue, voir les ch. 10 des vol. I et II.]

WEIZENBAUM, J. (1976) : *Computer power and human reason : from judgment to calculation*, W.H. Freeman, San Francisco.

WINOGRAD, T. (1983) : *Language as a cognitive process*, Volume 1 syntax, Addison Wesley.

2. Articles et recueils d'articles spécialisés :

[Les **ouvrages** dont les références suivent sont, pour le premier, une compilation commode où l'on peut trouver certaines des références mentionnées plus bas, et pour les suivants des publications collectives ou des actes de congrès consacrés notamment aux problèmes de dialogue.]

WEBBER, B.L. *& al.* (1986) : *Readings in Natural Language Processing,* San Mateo, Morgan Kaufmann.

REILLY, (1987) : *Communication failure in dialogue and discourse*, North-Holland Publishing Company.

TAYLOR, M., *& al.* (1989) : *Structure of multimodal dialog including voice*, North-Holland Publishing Company.

[un tome 2 est en préparation.]

COHEN, P. *& al.* (1990) : *Intentions in communication*, Bradford books at MIT Press.

NEEL, F. *& al.* (1992) : *Le dialogue homme-machine*, Actes du colloque, GRECO-PRC, Dourdan (à paraître).

[En ce qui concerne les **articles** dont les références suivent, l'article de H. P. Grice présente les "maximes" de la conversation, celui de D. Bobrow *& al.* le système GUS, celui de H. Sacks *& al.* les théories conversationalistes et celui de M. Guyomard *&al.* le seul système de dialogue oral actuellement utilisé, alors que les autres articles constituent les contributions les plus marquantes des années 80 en matière de signification intentionnelle.]

ALLEN, J. & PERRAULT, C.R. (1980) : Analyzing intention in utterances, *Artificial Intelligence* 15, 143-178 - repris dans B.L. Webber *& al.* (ed.) (1986).

BEROULE, D. & NEEL, F. (1984) : Une approche de problèmes liés à la communication parlée homme-machine, *Actes du 4ème congrès AFCET,* Paris, 345-354.

BOBROW, D.*& al.* (1077) : GUS, a frame-driven dialog system, *Artificial Intelligence*, 8, Amsterdam, Elsevier, 155-174 - repris dans B.L. Webber *& al.* (ed.) (1986).

CARBERRY, S. (1988) : Modelling the user's plans and goals, *Computational Linguistics*, 14, Allen, 23-37.

GRICE, H.P. (1979) : Logique et conversation, *Communications*, 30, Paris, Seuil, 57-72.

GROSZ, B. (1981) : Focusing and description in natural language dialogs, dans Webber, B. *& al.* (eds.) : *Discourse understanding*, Cambridge University Press, 85-105.

GROSZ, B. & SIDNER, C. (1986) : Attention, intentions, and the structure of discourse, *Computational linguistics*, 12, Allen .

GUYOMARD, M. *& al.* (1990) : Le rôle du dialogueur pour la reconnaissance de la parole. Le cas du système pages jaunes, *18ème JEP*, Montréal 322-326.

SACKS, H, *& al.* (1974) : A simplest systematics for the organisation of turn-taking in conversation, *Language*, 30 : 4, 696-735.

n° 509 du *Monde informatique* du 13.07.92 : "Homme-machine : un dialogue à réinventer"

3. Thèses françaises :

[Les quelques thèses qui suivent présentent l'avantage d'être écrites en français, d'inclure à chaque fois une présentation du domaine, et d'être assez faciles à trouver. A. Vilnat et M. Joab sont des thèses d'informatique, M. Joab traitant plus spécifiquement d'EIAO, P. Falzon est issu d'une thèse d'ergonomie, et D. Luzzati aborde la question d'un point de vue linguistique.]

FALZON, P. (1989) : *Ergonomie cognitive du dialogue*, Presses universitaires de Grenoble.

JOAB, M. (1990) : *Modélisation d'un dialogue pédagogique en langage naturel*, thèse d'Université, Paris VI.

LUZZATI, D. (1989) : *Recherches sur le dialogue homme-machine : modèles linguistiques et traitements automatiques*, thèse d'Etat, Paris III.

VILNAT, A. (1984) : *L'élaboration d'interventions pertinentes dans une conversation homme-machine*, thèse de 3ème cycle, Paris VI.

Annexes

INDEX DES TERMES

D

E

F

G

M

N

O

P

Q

R

S

T

U

V

INDEX DES NOMS PROPRES

M

N

O

P

Q

R

S

INDEX DES SIGLES

(NB: les sigles listés ci-dessous concernent des noms de systèmes, de théories ou de programmes scientifiques ; les renvois permettent de retrouver les unités dans l'index des termes)

Imprimé en France par I.M.E. — 25110 Baume-les-Dames
Dépôt légal n° 5111-06/1993
Collection n° 12 — Edition n° 01
14/4686/3

HU

Linguistique

Linguistique et Traitements Automatiques des Langues

Concevoir des logiciels capables de traiter des mots, des énoncés ou des textes de la langue, afin de dialoguer avec un utilisateur, d'aider l'humain à construire un texte, ou à le traduire : tel est l'objectif des traitements automatiques des langues. Sur ce domaine en pleine expansion, la présente **initiation**, destinée aux non-spécialistes et aux étudiants débutants, se propose de faire le point, en présentant les problématiques des traitements de la langue orale et écrite (théories linguistiques, formalismes de représentation et techniques informatiques), les limitations des réalisations actuelles, et les perspectives de recherche.

L'ouvrage se compose de deux parties. La première est consacrée aux **niveaux de traitement** de la langue : phonétique et phonologie, prosodie, morphologie, syntaxe, sémantique. La seconde porte sur les **domaines** des traitements automatiques : parole, traduction, compréhension, génération, dialogue homme-machine. Chaque chapitre est suivi de **repères bibliographiques** commentés permettant au lecteur d'approfondir ses connaissances.

Catherine FUCHS, chercheur au CNRS, dirige le Laboratoire de Linguistique ELSAP de Caen (associé au CNRS), et travaille sur le traitement formel de la sémantique.
Ses collaborateurs : **Anne LACHERET-DUJOUR**, enseignante à l'Université de Caen, spécialiste du traitement de la parole (LIMSI, Orsay) ; **Bernard VICTORRI**, chercheur au CNRS, spécialiste de modélisation sémantique et de réseaux connexionnistes (ELSAP, Caen) ; **Laurence DANLOS**, enseignante à l'Université Paris-VII, spécialiste de génération automatique de texte et de traduction automatique (TALANA, Paris-VII) ; **Daniel LUZZATI**, enseignant à l'Université du Mans, spécialiste du dialogue (LIMSI, Orsay).

Version Originale

9 782010 169083

14/4686/3
Imprimé en France
S.S.O.I. - PARIS